SUPERSTETH DIAGNOSTIC PROGRAMS, ROUTINES, AND STATISTICS (1951)

~~TOP SECRET CODEWORD~~

ACC NO

LOCATION

35751
CB0856
H01-0507-2
200-03

XHR

Box # H01-0507-2
00000047090004
Folder # 4
Paper

AHA Ø29
(3)

DECLASSIFIED
AUTHORITY NW 55104 19
BY NARA DATE: 4/25

Declassified by NSA/CSS
Deputy Associate Director for Policy and Records
On 20140218 by SG

~~TOP SECRET CODEWORD~~

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

~~SECRET~~

PROD-03

INFORMAL NO. 18
25 August 1957

SUBJECT: THE SUPERSTETH DIAGNOSTIC PROGRAMS

PREPARED BY: Paul D. Oyer

With an Appendix by: Edwin J. Hughes

DISTRIBUTION:

D/D	1	ADVA-43	2
SPEC. ASS'T.	1	ADVA-44	2
BLO	13	ACOM	1
USLO	5	ACOM-03	1
PROD	1	ACOM-1	2
PROD-03	315 (reserve)	ACOM-2	2
MPRO	1	ACOM-22	2
MPRO-03	6	ACOM-23	3
MPRO-1	13	ACOM-24	1
MPRO-3	11	ACOM-4	3
MPRO-5	18	ALLO	1
CREF	10	ALLO-03	1
CREF-407	1	ALLO-1	5
COLL-03	1	ALLO-2	5
GENS	1	ALLO-303	2
GENS-04	1	ALLO-31	1
GENS-14	2	ALLO-32	1
GENS-22	2	ALLO-33	2
GENS-35	2	ALLO-4	1
GENS-403	1	ALLO-403	2
GENS-43	4	ALLO-41	3
GENS-503	5	ALLO-42	1
ADVA	1	R/D	1
ADVA-03	1	ANEQ	6
ADVA-1	1	ENGR	1
ADVA-12	1	REMP	1
ADVA-13	2	REMP-1	2
ADVA-14	2	REMP-102	2
ADVA-2	1	DOC-2	1
ADVA-21	2	DOC-202	2
ADVA-22	3	STED-4	2
ADVA-23	1	TNG	1
ADVA-31	3	TNG-03	2
ADVA-4	1	TNG-2	1
		TNG-3	1

Declassified by NSA/CSS
Deputy Associate Director for Policy and Records
On 20140218 by SG

~~SECRET~~

DECLASSIFIED
NW5564S
AUTHORITY
BY NARA DATE: 4/25

SECRET

C O N T E N T S

PAGE	
1	The nature of the SUPERSTETH programs (Diagnostic tests and derived streams).
2	Summary of the programs (routines) available.
5	Description of the routines with statistics and examples.
5	Monographic Tests.
11	Digraphic Tests.
20	Monographic Local Roughness.
21	Polygraphic Repeats.
23	Width Tests.
25	Derived Streams (delta, decimation, Fibonacci).
27	Teleprinter Stream manipulations.
28	Substitutions.
30	Combined streams.
31	Additions to SUPERSTETH.
32	Explanation and examples of the use and interpretation of the Delta I.C. statistic and its sigmage. (Monographs, digraphs, ... , polygraphs).
44	The Chi-Squared Statistic (explanation and examples).
64	TABLES
65	Delta I.C. Table
66	Normal Table
67	Chi-Squared Table
68	Algebraic simplification of some diagnostic formulas.
72	Bibliography.

SECRET

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

SECRET

SUPERSTETH

INTRODUCTION: SUPERSTETH is a collection of computer programs which is used to perform selected diagnostic tests on a stream of input data or on some stream generated from it. For example, SUPERSTETH can be used to take monographic and digraphic frequency counts on an input stream, write out the stream on various widths, evaluating a delta IC for each column, generate a delta stream and make various statistical tests on this delta stream.

SUPERSTETH consists of a library of computer routines which make various counts, evaluate certain statistics, or generate certain new data streams. A sponsor can select from the library of routines those which make the desired tests and generate the desired streams. This selection is made by the use of the SUPERSTETH Rebus, which lists all available routines and, where appropriate, parameters that must be specified and thresholds that may be altered. For the convenience of the sponsor, there are a set of completed Rebus schedules, called Prescriptions, designed as standard SUPERSTETH initial attacks on various types of systems.

The input stream to SUPERSTETH may be a cipher or key stream of up to 12,000 characters in length, made up of any number of distinct alphabetic or numeric characters from 2 to 64, one of which may represent a dit. The sponsor may, if he so desires, specify a 5 or 6-level Baudot character coding to be used with his stream.

Sponsors are invited to obtain assistance from PROD-03 or their MPRO representative in requesting SUPERSTETH runs and in interpretation of results from both a statistical and cryptanalytic point of view.

NATURE OF DIAGNOSTIC TESTS: Diagnostic tests are applied to a stream of data to observe the statistical properties of the stream, with the object of determining the method of generation, arguing from effect to cause. In general, a diagnostic test consists of two computations

1. The evaluation of some statistic, or measure of a certain property of the data stream.

2. The evaluation of a sigmage, or measure of the significance of the numerical value of the statistic.

Thus, one of the diagnostic tests discussed above consists of two computations

1. The evaluation of the delta IC for each column of the stream written out on a specified width, a measure of the roughness in the column.

2. The evaluation of the sigmage of the delta IC, a measure of the probability that random data would exhibit such roughness.

The delta IC and Chi-squared statistics are used extensively in diagnostic tests. In the Appendices, these two statistics are discussed and examples of their use are given.

GENERATED OR DERIVED STREAMS: Diagnostic tests can be performed by SUPERSTETH not only on the original data stream, but also on other streams derived from it. Thus, one might want to study the properties of the delta stream, or perhaps the properties of the stream consisting of, say, every 5th character

SECRET

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

of the original stream, or the properties of the stream consisting of level ℓ of an original 5-level teletype character stream. SUPERSTETH can be used to generate such streams as

1. A delta stream (at any offset). (Routines 8.1.1, 8.1.2).
2. Decimation streams. (Routines 8.2.1, 8.2.2).
3. Teleprinter arithmetic streams. (Routines 8.4.1, 8.4.2).
4. De-Fibonaccied streams. (Routine 8.3.1).
5. Substitution streams. (Routine 8.5.1).
6. Combined streams. (Routine 9).

and to perform selected diagnostic tests on these generated streams.

DIAGNOSTIC TESTS AVAILABLE:

Monographic distribution (given in serial order or in frequency order).

Statistics: Delta IC, sigmage; Chi-squared, sigmage.
(Routines 1.1.0, 1.1.1, 1.2.1, 1.2.2).

Dot and cross level distributions.

Statistic: Sigmage of (cross)-count.
(Routine 1.3).

Digraphic distributions for the entire stream (overall or on-cut or off-cut).

(Routines 2.1.1, 2.4.1).

List of highest frequency digraphs. (Poisson)

(Routine 2.4.2)

Entire digraphic distribution statistics.

Statistics: Delta IC, sigmage; Chi-squared, sigmage.
(Routines 2.1.2, 2.1.3).

Row totals distribution, column totals distribution.

Statistics: Chi-squared, sigmage.
(Routines 2.2.1, 2.2.2).

Row totals distribution vs. column totals distribution.

Statistics: Chi-squared, sigmage.
(Routine 2.2.3).

Row totals distribution vs. column totals distribution at slides.

Statistics: Chi-squared, sigmage.
(Routine 2.2.4).

Distributions on diagonals.

Statistics: Chi-squared, sigmage.
(Routine 2.3.4).

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

Distribution on main diagonal; distribution symmetrically above and below main diagonal.

Statistics: Chi-squared, sigmage for Rutting, Symmetry, Bias.
(Routines 2.3.1, 2.3.2, 2.3.3).

Digraphic distributions for digraphs assembled according to a given rule.

(Digraph within group; skip digraphs).
(Routines 2.5, 2.5.1.1; 2.6, 2.6.0.1, 2.6.1.1).
The same tests may be performed on these digraphic distributions as on the digraphic distributions for the entire stream (see above).

Writeout on widths.

Column distributions.

Statistics: Delta IC, sigmage for each column; average columnar delta IC, root mean square sigmage.
(Routines 5.1, 5.2).

Column distribution vs. column distribution.

Statistics: Chi-squared, sigmage.
(Routine 5.3).

Column distribution vs. column distribution at slides.

Statistics: Chi-squared, sigmage.
(Routine 5.4).

Counts on position-in-group. (Column distributions for writeout on width of group size).

(Routine 1.4.1).

Column distributions.

Statistics: Delta IC, sigmage for each column; Chi-squared, sigmage for each column; total Chi-squared, sigmage for all columns.
(Routines 1.4.2, 1.4.3).

Column distribution vs. column distribution.

Statistics: Chi-squared, sigmage.
(Routine 1.4.4).

Column distribution vs. column distribution at slides.

Statistics: Chi-squared, sigmage.
(Routine 1.4.5).

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

Monographic local roughness. (Coincidence count on stream offset against itself).

Statistic: Sigmage of coincidence count.
(Routines 3.1, 3.2, 3.3)

Polygraphic repeats. (Locating long repeats).

Statistics: Delta IC, sigmage for the n-graphic distributions.
Index of cohesion.
(Routines 4.1, 4.2, 4.3, 4.4).

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

1. MONOGRAPHIC TESTS

1.1.0 ROUTINE. Printout of monographic distribution in the alphabetical order.

DESCRIPTION. Prints the monographic distribution in the alphabetical order as specified by the sponsor, together with totals of the number of valid characters, and the number of dits. (In the Rebus, the sponsor may label the alphabet with Roman or Cyrillic characters, digits, or Baudot configurations, using an arbitrary alphabetical ordering. The count for each character will be identified with the sponsor's label).

1.1.1 ROUTINE. Printout of monographic distribution in frequency order.

DESCRIPTION. Prints the monographic distribution in frequency order, the most frequent character first, together with totals of the number of valid characters and the number of dits.

1.2.1 ROUTINE. Monographic delta IC and sigmage.

DESCRIPTION. Prints the monographic delta IC and its sigmage.

$$\delta = \frac{\sum_{i=1}^{c} \frac{f_i(f_i-1)}{2}}{\frac{N(N-1)}{2} \cdot \frac{1}{c}} = \frac{c}{N(N-1)}\left[\left(\sum_{i=1}^{c} f_i^2\right) - N\right].$$

$$S(\delta) = \frac{\delta - 1}{\frac{1}{N}\cdot\sqrt{2(c-1)}} = \frac{N(\delta-1)}{\sqrt{2(c-1)}}.$$

δ = monographic delta IC. *

$S(\delta)$ = sigmage of the monographic delta IC. *

c = number of categories, or alphabet size.

f_i = number of valid characters in category i.

$N = \sum_{i=1}^{c} f_i$ = total number of valid characters.

* See Appendix A, pages 33-38 , for examples.

DECLASSIFIED
NW5561149
AUTHORITY
BY LJ NARA DATE: 4/25

1.2.2 ROUTINE. Chi-squared statistic on monographic distribution vs. a uniformly flat distribution. Sigmage of Chi-squared.

DESCRIPTION. Prints the Chi-squared statistic, X^2, comparing the monographic distribution against a uniformly flat distribution and the sigmage $S(\sqrt{2X^2})$.

$$X^2 = \sum_{i=1}^{c} \frac{\left(f_i - \frac{N}{c}\right)^2}{\frac{N}{c}} = \left(\frac{c}{N}\sum_{i=1}^{c} f_i^2\right) - N$$

$$S(\sqrt{2X^2}) = \sqrt{2X^2} - \sqrt{2(c-1)-1},$$

where X^2 = Chi-squared. *

NOTE 1. For $c \geq 30$ the statistic $\sqrt{2X^2}$ is approximately distributed according to the Normal distribution, with mean $\sqrt{2\nu - 1}$ and variance 1, where ν represents the number of degrees of freedom. For $c \geq 10$ the probability associated with $\sqrt{2X^2}$, obtained from a table of the Normal distribution, is often sufficiently accurate for practical purposes.

NOTE 2. The Chi-squared statistic which compares a distribution against a uniformly flat distribution is equivalent to the delta IC statistic, δ, and is related to it by the formula

$$X^2 = (N-1)(\delta - 1) + (c-1).$$

1.3.0 ROUTINE. Level counts, dot and cross distribution on individual levels.

DESCRIPTION. For each level, prints the distribution of dots and crosses, together with the statistic: sigmage of X(cross) count.

$$S_\ell = \frac{(m_\ell(X) - N P_\ell)}{\sqrt{N P_\ell (1 - P_\ell)}}$$

S_ℓ = sigmage of cross count on level ℓ.

$m_\ell(X)$ = number of crosses on level ℓ.

P_ℓ = probability of a cross on level ℓ. (This P_ℓ need not be $\frac{1}{2}$ as, for example, in a 26-letter, 5-level alphabet).

* See Appendix B , Example 1, pages 46-57.

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

1.4 MONOGRAPHS BY POSITION-IN-GROUP

1.4.1 ROUTINE. Monographic distribution in each column for stream written on width equal to group size.

PARAMETERS. Must specify group size.

DESCRIPTION. For selected group size 2 to 6, considering the stream as written on a width equal to the group size, counts and prints out the monographic distribution in each position of the group (in each column).

1.4.2 ROUTINE. Delta IC, sigmage; Chi-squared, sigmage statistics in each column for stream written on width equal to group size.

PARAMETERS. Must specify group size.

DESCRIPTION. For the distribution in each position of the group (for selected group size 2 to 6), prints the delta IC, sigmage of delta IC (using as expected value the overall delta IC, δ , rather than the flat random delta IC, 1), the Chi-squared statistic, χ^2_r , comparing the column distribution against that expected from row and column sums of a g-wide, c-deep contingency table (i.e. against that expected from the overall monographic distribution). Prints the sigmage, $S(\sqrt{2\chi^2_r})$.

$$\delta_r = \frac{\sum_{i=1}^{c} \frac{f_{ir}(f_{ir}-1)}{2}}{\frac{N_r(N_r-1)}{2} \cdot \frac{1}{c}} = \frac{c}{N_r(N_r-1)}\left[\left(\sum_{i=1}^{c} f_{ir}^2\right) - N_r\right],$$

$$r = 1, 2, \ldots, g.$$

$$S(\delta_r) = \frac{\delta_r - \delta}{\frac{1}{N_r}\cdot\sqrt{2(c-1)}} = \frac{N_r(\delta_r - \delta)}{\sqrt{2(c-1)}}.$$

$$\chi^2_r = \sum_{i=1}^{c} \frac{\left(f_{ir} - \frac{f_i N_r}{N}\right)^2}{\frac{f_i N_r}{N}} = \left[\frac{N}{N_r}\left(\sum_{i=1}^{c} \frac{f_{ir}^2}{f_i}\right)\right] - N_r.$$

$$S(\sqrt{2\chi^2_r}) = \sqrt{2\chi^2_r} - \sqrt{2\frac{(g-1)(c-1)}{g} - 1}.$$

DECLASSIFIED
NW55645
AUTHORITY
BY LJ NARA DATE: 4/25

where δ_r = monographic delta IC in column r .

X^2_r = Chi-squared, column r .

f_{ir} = number of valid characters in category i, column r .

$N_r = \sum_{i=1}^{c} f_{ir}$ = number of valid characters in column r .

$f_i = \sum_{r=1}^{g} f_{ir}$ = number of valid characters in category i.

g = group size.

NOTE. Here the statistic X^2_r, for a single column, is assumed to be distributed according to the Chi-squared distribution, with $\frac{(g-1)(c-1)}{g}$ degrees of freedom.

1.4.3 ROUTINE. Total Chi-squared (for all columns) for contingency table of position-in-group. Sigmage.

PARAMETERS. Must specify group size.

DESCRIPTION. For the selected group size, 2 to 6, prints out the total Chi-squared, $\hat{X}^2$, and the sigmage $S(\sqrt{2\hat{X}^2})$.

$$\hat{X}^2 = \sum_{r=1}^{g} X^2_r .$$

$$S(\sqrt{2\hat{X}^2}) = \sqrt{2\hat{X}^2} - \sqrt{2(g-1)(c-1) - 1} ,$$

where $\hat{X}^2$ = total Chi-squared, the sum of the Chi-squared scores for all the columns.

1.4.4 ROUTINE. Chi-squared on column vs. column for stream written on width equal to group size. Sigmage.

PARAMETERS. Must specify group size.

DESCRIPTION. For all combinations of pairs of positions of the selected group size 2 to 6, prints out the Chi-squared statistic for the 2 × c contingency table comparing the distributions of the pair of positions, and prints the sigmage.

$$X^2_{rs} = \sum_{i=1}^{c} \frac{\left(f_{ir} - \frac{N_r(f_{ir}+f_{is})}{N_{rs}}\right)^2}{\frac{N_r(f_{ir}+f_{is})}{N_{rs}}} + \sum_{i=1}^{c} \frac{\left(f_{is} - \frac{N_s(f_{ir}+f_{is})}{N_{rs}}\right)^2}{\frac{N_s(f_{ir}+f_{is})}{N_{rs}}}$$

DECLASSIFIED
NW5564S
AUTHORITY
BY LJ NARA DATE: 4/25

$$X^2_{rs} = \frac{N^2_{rs}}{N_r \cdot N_s}\left[\left(\sum_{i=1}^{c}\frac{f^2_{ir}}{(f_{ir}+f_{is})}\right) - \frac{N_r^2}{N_{rs}}\right], \quad r,s = 1,2,\ldots,8;\ r<s.$$

$$S(\sqrt{2X^2_{rs}}) = \sqrt{2X^2_{rs}} - \sqrt{2(c-1)-1},$$

Where X^2_{rs} = Chi-squared, column r vs. column s.

$N_{rs} = N_r + N_s$ = total valid characters in columns r and s.

See Appendix B, page 60 , Example 3 (where it is assumed that $N_r = N_s$)

1.4.5 ROUTINE. Chi-squared on column vs. column at slides for stream written on width equal to group size. Sigmage.

PARAMETERS. Must specify group size, slides.

DESCRIPTION. For each slide selected (from 1 to (c-1)), for all combinations of pairs of positions of the group size, prints out the Chi-squared statistic for the 2 × c contingency table comparing the distributions of the pair of positions at the selected slide, and prints the sigmage. (Note that slide ϕ is given in routine 1.4.4.)

$$_jX^2_{rs} = \sum_{i=1}^{c}\frac{\left(f_{ir} - \frac{(f_{ir}+f_{(i+j) \bmod c,s})}{2}\right)^2}{\frac{(f_{ir}+f_{(i+j) \bmod c,s})}{2}} + \sum_{i=1}^{c}\frac{\left(f_{(i+j) \bmod c,s} - \frac{(f_{ir}+f_{(i+j) \bmod c,s})}{2}\right)^2}{\frac{(f_{ir}+f_{(i+j) \bmod c,s})}{2}}.$$

$$_jX^2_{rs} = \sum_{i=1}^{c}\left[\frac{(f_{ir} - f_{(i+j) \bmod c,s})^2}{f_{ir}+f_{(i+j) \bmod c,s}}\right].$$

$$S(\sqrt{2\,{}_jX^2_{rs}}) = \sqrt{2\,{}_jX^2_{rs}} - \sqrt{2(c-1)-1}.$$

DECLASSIFIED
NW556449
AUTHORITY
BY LJ NARA DATE: 4/25

where $_{j}\chi^2_{rs}$ = Chi-squared, column r vs. column s at slide j.

$f_{(i+j) \bmod c, s}$ = number of characters in category k, column s, where $k = (i+j) \bmod c$ (Categories 0,1,2,..., 9 for digital text, c = 10, are selected, at slide j = 3, in the order 3,4,5,6,7,8,9,0,1,2).

See Example 3, Appendix B.

NOTE. The Chi-squared of 1.4.5 is a special case of that of 1.4.4. For simplification in 1.4.5 it is assumed that $N_r = N_s$.

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

2. DIGRAPHIC TESTS

2.1.1 ROUTINE. Printout of the digraphic distribution for overall or on-cut or off-cut count of the entire stream, with row and column totals.

DESCRIPTION. Prints out the selected overall, on-cut or off-cut digraphic distributions of the stream, showing row and column totals.

NOTE. In the stream ABCDEFGHI..., the overall count would include AB, BC, CD, DE, EF, ... etc; the on-cut count AB, CD, EF, GH, ... etc.; and the off-cut BC, DE, FG, HI.

2.1.2 ROUTINE. Delta IC statistic and sigmage on the selected (overall or on-cut or off-cut) digraphic distributions of the entire stream.

THRESHOLDS. May set threshold on the sigmage. If none specified, built-in threshold is 3.ϕ.

DESCRIPTION. Prints the delta IC statistic and its sigmage for the selected over-all or on-cut or off-cut digraphic distribution of the entire stream. If the sigmage exceeds 3.ϕ (or some specified value), prints the distribution (if it has not already been printed).

$$\delta_2 = \frac{\sum_{i=1}^{c}\sum_{j=1}^{c}\frac{f_{ij}(f_{ij}-1)}{2}}{\frac{M(M-1)}{2}\cdot\frac{1}{c^2}} = \frac{c^2}{M(M-1)}\left[\left(\sum_{i=1}^{c}\sum_{j=1}^{c} f_{ij}^2\right) - M\right].$$

$$S(\delta_2) = \frac{\delta_2 - 1}{\frac{1}{M}\sqrt{2\left[(c+1)^2 - 1\right]}} = \frac{M(\delta_2 - 1)}{\sqrt{2c(c+2)}}, \text{ for overall.}$$

$$S(\delta_2) = \frac{\delta_2 - 1}{\frac{1}{M}\sqrt{2\left[c^2 - 1\right]}} = \frac{M(\delta_2 - 1)}{\sqrt{2(c^2-1)}}, \text{ for on-cut, off-cut.}$$

DECLASSIFIED
NW5561649
AUTHORITY
BY LJ NARA DATE: 4/25

where δ_2 = digraphic delta IC.

M = total number of valid digraphs in the distribution considered.

f_{ij} = number of valid digraphs whose 1st character is in category i and 2nd in category j. $(i, j = 1, 2, \ldots, c)$.

See Appendix A, page 38, for an example.

2.1.3 ROUTINE. From the selected digraphic distribution, chi-squared statistic for the c x c contingency table and its sigmage.

THRESHOLDS. May set threshold on the sigmage. If none specified, built-in threshold is 3.ø.

DESCRIPTION. Prints out the chi-squared statistic and its sigmage on comparing the selected (overall, or on-cut, or off-cut) digraphic distribution against that expected from row and column sums, i.e. against that expected from the monographic distributions (not against flat random). If the sigmage exceeds 3.ø (or some specified value), prints the distribution (if it has not already been printed).

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{c} \frac{\left(f_{ij} - \frac{r_i s_j}{M}\right)^2}{\frac{r_i s_j}{M}} = M\left[\left(\sum_{i=1}^{c}\sum_{j=1}^{c} \frac{f_{ij}^2}{r_i s_j}\right) - 1\right]$$

$$S(\sqrt{2\chi^2}) = \sqrt{2\chi^2} - \sqrt{2(c-1)^2 - 1},$$

where $r_i = \sum_{j=1}^{c} f_{ij}$ = row sum = total valid characters in row i.

$s_j = \sum_{i=1}^{c} f_{ij}$ = column sum = total valid characters in column j.

See Appendix B, page 57, Example 2.

2.2.1 ROUTINE. From the selected digraphic distribution, chi-squared statistic for distribution of row sums vs. a uniformly flat distribution. Sigmage of chi-squared.

DESCRIPTION. Prints the chi-squared statistic and its sigmage in comparing the row sums distribution against a uniformly flat distribution.

$$\chi_r^2 = \sum_{i=1}^{c} \frac{\left(r_i - \frac{M}{c}\right)^2}{\frac{M}{c}} = \left(\frac{c}{M}\sum_{i=1}^{c} r_i^2\right) - M.$$

$$S(\sqrt{2\chi_r^2}) = \sqrt{2\chi_r^2} - \sqrt{2(c-1) - 1}.$$

See Appendix B, page 54, Example 1.

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

2.2.2 ROUTINE. From the selected digraphic distribution, chi-squared statistic for the distribution of column sums vs. a uniformly flat distribution. Sigmage of chi-squared.

DESCRIPTION. Same as 2.2.1, with columns substituted for rows.

2.2.3 ROUTINE. From the selected digraphic distribution, chi-squared statistic for row sums distribution vs. column sums distribution. Sigmage of chi-squared.

DESCRIPTION. Prints the chi-squared statistic for the 2 x c contingency table comparing the distributions of the row sums and column sums, and prints the sigmage.

$$\chi^2 = \sum_{i=1}^{c} \frac{\left(r_i - \frac{(r_i + s_i)}{2}\right)^2}{\frac{(r_i + s_i)}{2}} + \sum_{i=1}^{c} \frac{\left(s_i - \frac{(r_i + s_i)}{2}\right)^2}{\frac{(r_i + s_i)}{2}}$$

$$\chi^2 = \sum_{i=1}^{c} \frac{(r_i - s_i)^2}{(r_i + s_i)}.$$

$$S(\sqrt{2\chi^2}) = \sqrt{2\chi^2} - \sqrt{2(c-1)-1}\ .$$

See Appendix B, page 60, Example 3.

2.2.4 ROUTINE. From the selected digraphic distribution, chi-squared statistic for row sums distribution vs. column sums distribution at selected slides.

PARAMETERS. Must specify slides.

DESCRIPTION. For each slide selected (from 1 to (c-1)), prints the chi-squared statistic for the 2 x c contingency table comparing the distributions of the row sums and column sums at the given slide, and prints the sigmage.

$${}_j\chi^2 = \sum_{i=1}^{c} \frac{\left(r_i - \frac{(r_i + s_{(i+j) \bmod c})}{2}\right)^2}{\frac{r_i + s_{(i+j) \bmod c}}{2}} + \sum_{i=1}^{c} \frac{\left(s_{(i+j) \bmod c} - \frac{(r_i + s_{(i+j) \bmod c})}{2}\right)^2}{\frac{r_i + s_{(i+j) \bmod c}}{2}}$$

$${}_j\chi^2 = \sum_{i=0}^{c-1} \frac{(r_i - s_{(i+j) \bmod c})^2}{(r_i + s_{(i+j) \bmod c})}, \quad \text{for } j = 1, 2, \ldots, (c-1).$$

$$S(\sqrt{2\,{}_j\chi^2}) = \sqrt{2\,{}_j\chi^2} - \sqrt{2(c-1)-1}\ .$$

DECLASSIFIED.
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

where $S_{(i+j) \bmod c} = S_k$ = column sum = total valid characters in column k, where $k = (i+j) \bmod c$. (Sums of columns k for digital text, c = 10, are selected, at slide j = 3, in the order k = 3,4,5,6,7,8,9,0,1,2).

See Example 3, Appendix B.

2.3 DIAGONAL AND SYMMETRY STATISTICS

2.3.1 ROUTINE. From the selected digraphic distribution, chi-squared statistic for rutting, i.e. chi-squared for the main diagonal (doublet distribution). Sigmage of chi-squared.

DESCRIPTION. Prints the chi-squared statistic and its sigmage in comparing the distribution for the main diagonal against that which would be expected from row and column sums.

$$X^2_{01} = \sum_{i=1}^{c} \frac{\left(f_{ii} - \frac{r_i s_i}{M}\right)^2}{\frac{r_i s_i}{M}} = \frac{1}{M} \sum_{i=1}^{c} \frac{\left(M f_{ii} - r_i s_i\right)^2}{r_i s_i}.$$

$$S\left(\sqrt{2X^2_{01}}\right) = \sqrt{2X^2_{01}} - \sqrt{2c-1}.$$

Where X^2_{01} = chi-squared for rutting.

NOTE. It is assumed that this statistic is distributed according to the chi-squared distribution, and that the number of degrees of freedom is c. Example: In the dinomic matrix, page 59, the main diagonal frequencies, f_{ii} $(i = 0,1,2,\ldots,9)$, are 0,18,0,7,2,8,1,5,1,2. $(c = 10)$

2.3.2 ROUTINE. From the selected digraphic distribution, chi-squared statistic for symmetry, i.e. chi-squared for lower left triangle summed to upper right triangle of the contingency table. Sigmage of chi-squared.

DESCRIPTION. Prints the chi-squared statistic and its sigmage in comparing the distribution of the lower left triangle summed to the upper right triangle of the contingency table against that expected from row and column sums.

$$X^2_{02} = \sum_{\substack{i=1 \\ (i<j)}}^{c} \sum_{j=1}^{c} \frac{\left[(f_{ij} + f_{ji}) - \frac{(r_i s_j + r_j s_i)}{M}\right]^2}{\frac{(r_i s_j + r_j s_i)}{M}}.$$

$$X^2_{02} = \frac{1}{M} \sum_{\substack{i=1 \\ (i<j)}}^{c} \sum_{j=1}^{c} \frac{\left[M(f_{ij} + f_{ji}) - (r_i s_j + r_j s_i)\right]^2}{(r_i s_j + r_j s_i)}$$

$$S\left(\sqrt{2X^2_{02}}\right) = \sqrt{2X^2_{02}} - \sqrt{2\left[\frac{(c-1)(c-2)}{2}\right] - 1} = \sqrt{2X^2_{02}} - \sqrt{(c-1)(c-2) - 1}.$$

where X^2_{02} = chi-squared for symmetry.

NOTE. It is assumed that this statistic is distributed according to the chi-squared distribution, and that the number of degrees of freedom is $1 + 2 + \ldots + (c-2) = \frac{1}{2}(c-1)(c-2)$

DECLASSIFIED
NW5561149
AUTHORITY
BY LJ NARA DATE: 4/25

2.3.3 ROUTINE. From the selected digraphic distribution, chi-squared statistic for bias, equal to the χ^2 of 2.1.3 minus chi-squared for rutting and symmetry. Sigmage of chi-squared.

DESCRIPTION. Prints the chi-squared statistic, and its sigmage, obtained by subtracting from the total chi-squared statistic the chi-squared statistics for rutting and symmetry.

$$\chi^2_{03} = \chi^2 - \chi^2_{01} - \chi^2_{02}$$

$$\chi^2_{02} = M \sum_{\substack{i=1 \\ (i<j)}}^{c} \sum_{j=1}^{c} \frac{(f_{ij} - f_{ji})^2}{r_i s_j + r_j s_i}$$

$$S(\sqrt{2\chi^2_{03}}) = \sqrt{2\chi^2_{03}} - \sqrt{2\left[\frac{c(c-3)}{2}\right] - 1}$$

$$S(\sqrt{2\chi^2_{03}}) = \sqrt{2\chi^2_{03}} - \sqrt{c(c-3) - 1},$$

where χ^2_{03} = chi-squared for bias.

NOTE. The number of degrees of freedom is $(c-1)^2 - c - \frac{(c-1)(c-2)}{2} = \frac{c(c-3)}{2}$

$$= \frac{(c-1)(c-2)}{2} - 1.$$

2.3.4 ROUTINE. From the selected digraphic distribution, chi-squared statistic for comparing the distribution along a diagonal against that expected from row and column sums. Sigmage of chi-squared.

DESCRIPTION. Prints the chi-squared statistic and its sigmage in comparing the distribution along the jth diagonal against that which would be expected from row and column sums, for each $j = 0, 1, 2, \ldots, (c-1)$.

$$\chi^2_{D_j} = \sum_{i=1}^{c} \frac{\left(f_{i,(i+j) \bmod c} - \frac{r_i s_{(i+j) \bmod c}}{M}\right)^2}{\frac{r_i s_{(i+j) \bmod c}}{M}}$$

$$\chi^2_{D_j} = \frac{1}{M} \sum_{i=1}^{c} \left[\frac{\left(M f_{i,(i+j) \bmod c} - r_i s_{(i+j) \bmod c}\right)^2}{r_i s_{(i+j) \bmod c}}\right]$$

$$S(\sqrt{2\chi^2_{D_j}}) = \sqrt{2\chi^2_{D_j}} - \sqrt{2c-1}.$$

Note that when $j = \emptyset$ in routine 2.3.4. the statistics are for the main diagonal, as in 2.3.1.

Example:

For a digital distribution ($c = 10$), it is more convenient to run the index i, in the summation symbol, from $\emptyset$ to 9 rather than 1 to 10. Thus

$$\sum_{i=1}^{c} = \sum_{i=1}^{10} \quad \text{may be written} \quad \sum_{i=0}^{9}$$

DECLASSIFIED.
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

The symbol (i+j) mod c means that if (i+j) is equal to or larger than c, then c is subtracted from (i+j). The diagonals all run from northwest to southeast.

For j = 1, the diagonal D_1 consists of the dinomic frequencies (i = 0,1,2,...,9); $f_{\emptyset 1}$, f_{12}, f_{23}, f_{34}, f_{45}, f_{56}, f_{67}, f_{78}, f_{89}, and $f_{9\emptyset}$. In the dinomic matrix (labeled mononome-dinome) of Example 2, Appendix B, page 58, the frequencies in diagonal D_1 are $f_{01}=1$, $f_{12}=3$, $f_{23}=10$, $f_{34}=8$, $f_{45}=1$, $f_{56}=\emptyset$, $f_{67}=3$, $f_{78}=2$, $f_{89}=1$, and $f_{9\emptyset}=11$.

Then $\chi^2_{D_1}$ is computed as follows: ($M=323$, $r_0 = 46$, $s_1 = 33$, etc.)

$$\chi^2_{D_1} = \frac{\left(1 - \frac{46\times 33}{323}\right)^2}{\frac{46\times 33}{323}} + \frac{\left(3 - \frac{33\times 29}{323}\right)^2}{\frac{33\times 29}{323}} + \ldots + \frac{\left(11 - \frac{36\times 46}{323}\right)^2}{\frac{36\times 46}{323}}.$$

$$\chi^2_{D_1} = 2.91 + 0.00 + 9.27 + \ldots + 6.72 = 27.03.$$

This is a chi-squared with 10 degrees of freedom. The score, 27.03, may be interpreted by means of an appropriate chi-squared table. Alternately, the result obtained from the sigmage formula may be interpreted by means of a table of the normal distribution. The sigmage, in this case, is

$$S\left(\sqrt{2\chi^2_{D_1}}\right) = \sqrt{54.06} - \sqrt{19} = +2.99 \text{ sigmas.}$$

2.4.1 ROUTINE. Poisson print out of the selected digraphic distribution.

DESCRIPTION. For the selected overall or on-cut of off-cut digraphic distribution prints out the number of digraphs that occur each number of times: $\emptyset$, 1, 2, ..., (t-1), (t or more), where t is the smallest integer equal to or greater than the expected count per cell plus 10 sigmas.

$$t \geq \frac{M}{C^2} + 10\sqrt{M \times \frac{1}{C^2} \times \left(1 - \frac{1}{C^2}\right)}.$$

2.4.2 ROUTINE. From the selected digraphic distribution, print out of those digraphs that occur m or more times.

DESCRIPTION. For the slected digraphic distribution prints out the actual digraphs that occur m or more times, where m is equal to the expected count per cell plus 8 sigmas, together with the number of occurrences.

$$m \geq \frac{M}{C^2} + 8\sqrt{M \times \frac{1}{C^2} \times \left(1 - \frac{1}{C^2}\right)}.$$

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

2.5 ROUTINE. Formation of digraphs within group

PARAMETERS. Must specify group size.

DESCRIPTION. For the specified group size, counts the digraphs formed by the characters in the positions specified below. For each pair of positions there is a separate digraphic distribution.

Group size	Positions
2	1-2, 2-1.
3	1-2, 1-3, 2-3, 3-1.
4	1-2, 1-3, 1-4, 2-3, 2-4, 3-4, 4-1.
5	1-2, 1-3, 1-4, 1-5, 2-3, 2-4, 2-5, 3-4, 3-5, 4-5, 5-1.
6	1-2, 1-3, 1-4, 1-5, 1-6, 2-3, 2-4, 2-5, 2-6, 3-4, 3-5, 3-6, 4-5, 4-6, 5-6, 6-1.

NOTE. 2-1, 3-1, 4-1, etc. indicate the last character in one group and the first character in the following group.

This routine makes available the above digraphic distributions for statistical analysis by routines selected from the list:

Routines 2.5.1.1, 2.1.2, 2.1.3, 2.2.1, 2.2.2, 2.2.3, 2.3.1, 2.3.2, 2.3.3, 2.3.4, and 2.4.1.

2.5.1.1. ROUTINE.. Print out of digraphic distributions for digraphs within group.

PARAMETERS. Must specify group size.

DESCRIPTION. Prints out the digraphic distributions specified in routine 2.5.

2.6 ROUTINE. Formation of digraphs as specified by sponsor.

PARAMETERS. Must specify 3 parameters i,j,k.

DESCRIPTION. For each set of selected values of the parameters i,j,k count the digraphs formed as follows: the first digraph is formed by the character of the stream in position i as first element and by that in position j as second element. The parameter k, the skip, is added to both i and j successively to locate the positions of the elements of successive digraphs. This process of forming digraphs is continued until the stream is exhausted. The difference (j-i) is called the offset.

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

This routine forms such digraphic distributions and makes them available for statistical analysis by routines selected from the list:

Routines 2.6.1.1, 2.1.2, 2.1.3, 2.2.1, 2.2.2, 2.2.3, 2.3.1, 2.3.2,

2.3.3, 2.3.4, 2.4.1.

Example:

Stream: AJQD MTZL BAZO PGEU MCXH OYEN ...

For $i = 1$, $j = 5$, $k = 4$; the digraphs counted are

AM, MB, BP, PM, MO, ...

(This is the same as a vertical digraph count of column 1 at width 4).

For $i = 2$, $j = 6$, $k = 4$; the digraphs counted are

JT, TA, AG, GC, CY, ...

For $i = 3$, $j = 7$, $k = 4$;

QZ, ZZ, ZE, EX, XE, ...

For $i = 4$, $j = 8$, $k = 4$;

DL, LO, OU, UH, HN, ...

2.6.0.1 ROUTINE. Formation of combined digraphic distributions.

DESCRIPTION. Makes available in one digraphic matrix the sum of the counts for all distributions corresponding to selected sets of parameters i,j, and k of routine 2.6. The combined digraphic distribution is available for statistical analysis by routines selected from the list:

Routines 2.6.1.1, 2.1.2, 2.1.3, 2.2.1, 2.2.2, 2.2.3, 2.3.1, 2.3.2,

2.3.3, 2.3.4, 2.4.1.

Example:

For the example of 2.6, the combined distribution will consist of the digraphs

AM, MB, BP, PM, MO, ..., JT, TA, AG, GC, CY, ...,

QZ, ZZ, ZE, EX, XE, ..., DL, LO, OU, UH, HN, ...

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

2.6.1.1 ROUTINE. Printout of the digraphic distributions of 2.6 specified by the sponsor.

DESCRIPTION. Prints out the distributions.

DECLASSIFIED
NW55645
AUTHORITY
BY LJ NARA DATE: 4/25

3. MONOGRAPHIC LOCAL ROUGHNESS.

3.1 **ROUTINE.** Coincidence count on stream offset against itself.

PARAMETERS. Must specify offsets or range of offsets.

THRESHOLDS. May set threshold on the sigmage. If none specified, built in threshold set at 3.5.

DESCRIPTION. For each offset selected, of any size or in any range from one up to stream length minus c, prints the offset, the total number of monographic coincidences, the expected number of coincidences, and the sigmage. If any sigmage exceeds 3.5 (or other specified value) prints out the distribution of coincident letters for that offset.

$$E(H) = \frac{P\delta}{c}$$

$$S(H) = \frac{H - \frac{P\delta}{c}}{\sqrt{P\left(\frac{\delta}{c}\right)\left(1-\frac{\delta}{c}\right)}} = \frac{cH - P\delta}{\sqrt{P\delta(c-\delta)}},$$

where H = number of coincidences

P = number of valid comparisons

δ = monographic delta IC.

3.2 **ROUTINE.** Subtotals of monographic coincidence count according to "position-in-group."

PARAMETERS. Must specify group size.

DESCRIPTION. For the selected offset, equal to the group size, prints subtotals of monographic coincidences according to position in group.

EXAMPLE. For offset 3 and the stream ABCAFCDFEAJE the subtotals are

Position-in-group 1.: 1
2.: 1
3.: 2

DECLASSIFIED
NW5556449
AUTHORITY
BY LJ NARA DATE: 4/25

~~SECRET~~

4. POLYGRAPHIC REPEATS

4.1 ROUTINE. Delta IC and sigmage for n-graphic distributions.*

DESCRIPTION. For each size n-graph (polygraph n characters long) $n = 2,3,4, \ldots, P$, where P is chosen (as given below) such that the expected number of P-graphic repeats in the stream is less than 20, prints the total number of n-graphs, and the delta IC and sigmage of the n-graphic distribution.

P is the smallest integer in the set $n = 3,4,5,\ldots$, such that

$$\frac{N(N-1)}{2} \cdot \frac{1}{c^n} \leq 20 .$$

$$\delta_n = \frac{\sum_{i=1}^{c^n} \frac{f_i(f_i-1)}{2}}{\frac{N(N-1)}{2} \cdot \frac{1}{c^n}} = \frac{c^n}{N(N-1)} \left[\left(\sum_{i=1}^{c^n} f_i^2 \right) - N \right].$$

$$S(\delta_n) = \frac{(\delta_n - \delta_{n-1} \cdot \delta_1)}{\frac{1}{N}\sqrt{2(c^n-1)}} = \frac{N(\delta_n - \delta_{n-1} \cdot \delta_1)}{\sqrt{2(c^n-1)}} ,$$

where N = the total number of n-graphs $(n = 2,3,\ldots,P)$.*
δ_n = delta IC for the n-graphic distribution.
δ_1 = delta IC for the monographic distribution.
f_i = the frequency of the i-th distinct n-graph.*

4.2 ROUTINE. Index of Cohesion between (n-1) and n-graphic distributions.*

DESCRIPTION. For each size n-graph $(n = 2,3,\ldots, P)$ show the index of cohesion $i_n = \frac{\delta_n}{\delta_{n-1} \cdot \delta_1}$ where P is as in 4.1 and where i_n = index of cohesion between (n-1) and n-graphic distributions.

4.3 ROUTINE. Printout of repeats and the interval of spacing.

THRESHOLDS. May set threshold on sigmage. If none specified, built-in threshold is 3.5.

DESCRIPTION. For the P determined in 4.1, prints the following information for all repeated P-graph or longer polygraphs. For each repeat prints the position of the first character of its first occurrence and the offset to the first character of its second occurrence, also the length of the repeat and the repeat itself.

* See Appendix A, pages 33-35 and 40-43 for examples and explanations.

~~SECRET~~

DECLASSIFIED
NW5561+9
AUTHORITY
BY LJ NARA DATE: 4/25

If for any integer m in the sequence $m = 3,4,5,\ldots$, P the sigmage $S(\delta_m)$, computed in 4.1, is equal to or greater than 3.5, (or other specified value), then the routine prints the same information as in the previous paragraph for all repeated m-graphs or longer polygraphs.

NOTE. Prints polygraphs in their longest form. Thus, for example, a hexagraph in the print out is not considered as two consecutive pentagraphs, and HOGMJP is not printed as HOGMJ and OGMJP.

4.4 ROUTINE. Sum Check, modulo c, between repeats.

PARAMETERS. Must specify parameter k.

DESCRIPTION. Prints out, for each polygraphic repeat of size (P + k) or longer, the sum (modulo c) of the stream which starts with the first character of the first occurrence of the polygraph and ends with the last character before the second occurrence. Any dit as part of the sum shall yield a sum of dit. (The number k is a non-negative integer to be specified by the sponsor.).

Example. For the polygraphic repeat 523 shown below

... 07852357186452396l7 ...

the sum is $5+2+3+5+7+1+8+6+4 = 1 \pmod{10}$.

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

5. WIDTH WRITEOUTS.

5.1 ROUTINE. Average columnar delta IC on widths, sigmage.

PARAMETERS. Must specify widths, or range of widths.

DESCRIPTION. For each width in a range of widths selected (from 2 to 128) and for each other width selected (over 128), prints the width size, the average columnar delta IC and the root mean square sigmage.

$$ {}_w\bar{\delta} = \sum_{k=1}^{w} \frac{N_k \cdot {}_w\delta_k}{N} . $$

$$ \bar{S} = \sqrt{\frac{1}{w}\left(\sum_{k=1}^{w} [S({}_w\delta_k)]^2\right)} , $$

where ${}_w\bar{\delta}$ = average columnar delta IC

$\bar{S}$ = root mean square sigmage.

N_k = number of valid characters in column k.

${}_w\delta_k$ = delta IC for column k, width w.

Note: The expected value of $\bar{S}$ is 1.0 . The quantity $w \cdot (\bar{S})^2 = ({}_wS_1)^2 + ({}_wS_2)^2 + \cdots + ({}_wS_w)^2$ is approximately chi-squared with w degrees of freedom.

5.2 ROUTINE. Monographic delta IC and sigmage for each column of stream written on selected widths.

PARAMETERS. Must specify widths, or range of widths.

THRESHOLDS. Must set threshold on sigmage. If none specified, built-in threshold is 3.0.

DESCRIPTION. For each column of each width selected, computes the monographic delta IC and its sigmage. For those columns with sigmage greater than 3.0 (or other specified value), prints the width size, the column number, the delta IC, and its sigmage.

$$ {}_w\delta_k = \frac{\sum_{i=1}^{c} \frac{f_{ik}(f_{ik}-1)}{2}}{\frac{N_k(N_k-1)}{2} \cdot \frac{1}{c}} = \frac{c}{N_k(N_k-1)}\left[\left(\sum_{i=1}^{c} f_{ik}^2\right) - N_k\right] ; $$

for $k = 1, 2, \ldots, w$.

$$ {}_wS_k = S({}_w\delta_k) = \frac{{}_w\delta_k - \delta}{\frac{1}{N_k}\sqrt{2(c-1)}} = \frac{N_k({}_w\delta_k - \delta)}{\sqrt{2(c-1)}} , $$

DECLASSIFIED
NW556649
AUTHORITY
BY LJ NARA DATE: 4/25

where f_{ik} = frequency of category i in column k.

NOTE. The sigmage of a columnar delta IC is found using as expected value the overall monographic delta IC, δ, rather than the random expected value 1.

5.3 ROUTINE. Chi-squared statistic on column vs column from width writeout. Sigmage of Chi-squared.

PARAMETERS. Must specify identity of two columns and width.

DESCRIPTION. For each selected set of parameters i, j; w specifying a pair of columns, i, j, of the stream written on width w, compares the distributions of the columns by means of the Chi-squared and sigmage statistics of routine 1.4.4 and prints the parameters and these statistics.

5.4 ROUTINE. Chi-squared statistic on column vs. column at slides, from width writeout. Sigmage of Chi-squared.

PARAMETERS. Must specify slides, identity of two columns, and width.

DESCRIPTION. For each selected set of parameters s; i,j,; w specifying a pair of columns i,j of the stream written on width w, and a slide s to be applied to column j, compares the distribution of column i against that of column j with slide s by means of the Chi-squared and sigmage statistics of routine 1.4.5 and prints the parameters and these statistics.

DECLASSIFIED
NW5561+9
AUTHORITY
BY LJ NARA DATE: 4/25

8. DERIVED STREAMS

8.1.1 ROUTINE. Form the delta-n stream.

PARAMETERS. Must specify the delta interval n and the modulus of arithmetic.

DESCRIPTION. Forms the delta-n stream for a selected integer n $[1 \leq m \leq$ (input stream length minus c)$]$. The subtraction will be taken modulo c (or any other chosen modulus up to 64). If the stream is S_1, S_2, S_3, ..., S_N, then the delta-n stream is

$$(S_{m+1} - S_1),\ (S_{m+2} - S_2),\ \ldots,\ (S_N - S_{N-m}).$$

(ϕ is used for the first character of the alphabet, 1 for the second, ..., and (c-1) for the last character of the alphabet.)

NOTE. If the modulus is larger than c, symbols are assigned according to the list:

A, B, C, ..., Y, Z, 0, 1, ..., 9, 37-, 38-, ..., 64-.
(in any prescribed order).

EXAMPLE:

If the alphabet is of size c=26, but the modulus for subtraction is chosen as 32, then the delta-n stream will consist of the following characters:

A, B, C, ..., Z, 0, 1, 2, 3, 4, 5. (with this order prescribed)

Examples of subtraction when C=26 but modulus=32:

J - H=C, M - M=A, Q - R=5 (the 32nd character),

L - T=Y, P - S=3 (the 30th character).

8.1.2 ROUTINE. Form the delta-n stream, with subtraction reversed.

PARAMETERS. Must specify the delta interval n and the modulus of arithmetic.

DESCRIPTION. Same as 8.1.1 except that if the stream is S_1, S_2, S_3, ..., S_N, then the delta-n stream is

$$(S_1 - S_{m+1}),\ (S_2 - S_{m+2}),\ \ldots,\ (S_{N-m} - S_N).$$

DECLASSIFIED
NW5564S
AUTHORITY
BY NARA DATE: 4/25

SECRET

8.2.1 ROUTINE. Form decimation streams at an interval of n.

PARAMETERS. Must specify the interval of decimation, n.

DESCRIPTION. For decimation n, forms any selection of up to n decimation sub-streams ${}_i d_m$ ($i = 1, 2, \ldots, n$; $2 \leq m \leq (1/2) \cdot$stream length).

EXAMPLE: $n=3$

${}_1d_3 : S_1, S_4, S_7, S_{10}, \ldots$

${}_2d_3 : S_2, S_5, S_8, S_{11}, \ldots$

${}_3d_3 : S_3, S_6, S_9, S_{12}, \ldots$

8.2.2 ROUTINE. Combine decimation sub-streams.

PARAMETERS. Must specify the interval of decimation, n, and the order in which streams are to be recombined.

DESCRIPTION. Combines a selection of the n sub-streams in any prescribed order into one long stream.

EXAMPLE. Combining all 3 sub-streams of the previous example in 1, 3, 2 order would give the stream.

$S_1, S_4, S_7, S_{10}, \ldots, S_3, S_6, S_9, S_{12}, \ldots,$

$S_2, S_5, S_8, S_{11}, \ldots$.

8.3.1 ROUTINE. Form de-fibonaccied stream.

PARAMETERS. Must specify the parameters j and k for the Fibonacci generation.

DESCRIPTION. Forms, from an original stream $S_1, S_2, S_3, \ldots$, a de-fibonaccied stream using the following rule of combination, modulo c:

$$\lfloor (S_1 + S_j) - S_k \rfloor, \lfloor (S_2 + S_{j+1}) - S_{k+1} \rfloor, \lfloor (S_3 + S_{j+2}) - S_{k+2} \rfloor \ldots$$

The parameter j must be in the range $1 \leq j \leq 51$ and k: $2 \leq k \leq 101$, (always $j < k$).

SECRET

DECLASSIFIED
NW5516 49
AUTHORITY
BY LJ NARA DATE: 4/25

8.4.1 ROUTINE. Form Teleprinter Arithmetic Streams.

PARAMETERS. Must specify the "delta" interval n.

DESCRIPTION. Using sponsor's coding obtains "delta-n" streams at selected interval n $(1 \leq n \leq (N-c))$ from level by level combination of stream characters:

$$(S_1 , S_{M+1}), (S_2 , S_{M+2}), (S_3 , S_{M+3}), \ldots$$

under ordinary baud addition (rule 1, below) or under any of 10 rules of combination of 2 binary digits:

INPUT COMBINATIONS	RULES FOR OUTPUT	1	2	3	4	5	6	7	8	9	10
0,0		0	1	0	0	0	0	1	1	1	1
0,1		1	0	1	0	0	1	0	1	1	0
1,0		1	0	0	0	1	1	1	1	0	0
1,1		0	1	0	1	0	1	1	0	1	0

EXAMPLE:

Stream:

1 0 1 1 0 . . .
0 0 1 0 1 . . .
1 1 0 0 0 . . .
0 1 1 0 0 . . .
1 0 1 1 0 . . .

Using rule 1, namely

	0	1
0	0	1
1	1	0

, the delta-1 stream is

1 1 0 1 . . .
0 1 1 1 . . .
0 1 0 0 . . .
1 0 1 0 . . .
1 1 0 1 . . .

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

8.4.2 ROUTINE. Convert character stream into a single stream of bits.

DESCRIPTION. Converts the stream of characters into a single stream of bits by reading all levels of the first character, then all levels of the second character then all levels of each succeeding character. In every case, levels within a character are read in a sequence selected by the sponsor.

EXAMPLE. The stream of 6-level characters

```
0 1 1 . . .
1 0 1 . . .
0 0 1 . . .
0 1 0 . . .
1 1 0 . . .
0 0 0 . . .
```

read in 1 2 3 4 5 6 order becomes

0 1 0 0 1 0 1 0 0 1 1 0 1 1 1 0 0 0 . . .

and read in 3 5 2 1 4 6 order becomes

0 1 1 0 0 0 0 1 0 1 1 0 1 0 1 1 0 0 . . .

For streams of 5-level characters the derived stream of bits is 5 times as long as the input stream; for streams of 6-level characters the derived stream of bits is 6 times as long as the input stream.

8.4.3 ROUTINE. Shifted - levels teletype stream.

DESCRIPTION. Derived stream is formed from an input stream of 6-level characters by shifting or transposing the levels of each character in a prescribed pattern.

EXAMPLE. For pattern 1 2 3 4 5 6 shifted to 3 2 1 4 6 5

Levels	Input	Derived Output
1	0 1 1 . . .	0 0 1 . . .
2	1 0 1 . . .	1 0 1 . . .
3	0 0 1 . . .	0 1 1 . . .
4	0 1 0 . . .	0 1 0 . . .
5	1 1 0 . . .	0 0 0 . . .
6	0 0 0 . . .	1 1 0 . . .

8.5.1 ROUTINE. Form substitution stream.

DESCRIPTION. Derives a new stream by substitution from a given alphabet of an original stream to a new alphabet, using a selected type of substitution, as specified by the sponsor. A list of various types is included below.

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

~~SECRET~~

TYPE 1. Substitution n to 1 (where $1 < n < S/2$ for at least one character of the final alphabet).

EXAMPLE. Substitution table.

0	1	2	3	4	5	6	7	8	9
A	B	C	D	E	F	G	H	I	J
K	L	M	N	O	P	Q	R	S	T
U	V	W	X	Y	Z				

In the derived stream A, K, U become ∅; B, L, V become 1; etc.

TYPE 2. Substitution 1 for 1 or 2 for 1. (Monome-Dinome)

EXAMPLE. Substitution table.

	3	4	5	6	7	8	9
	A	B	C	D	E	F	G
0	H	I	J	K	L	M	N
1	O	P	Q	R	S	T	U
2	V	W	X	Y	Z		

In the derived stream A through G convert to mononomes (e.g. A becomes 3) and H through Z convert to dinomes (e.g. H becomes 03, Z becomes 27)

TYPE 3. Substitutions 2 for 1

EXAMPLE. Substitution table.

	5	6	7	8	9
0	A	B	C	D	E
1	F	G	H	I	K
2	L	M	N	O	P
3	Q	R	S	T	U
4	V	W	X	Y	Z

In the derived stream each letter converts to a dinome. A becomes 05, B becomes 06, ..., Z becomes 49.

~~SECRET~~

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

TYPE 4. Substitution 1 for 2.

EXAMPLE. Using the same table as in the previous example, 05 converts to A, 06 to B, ..., 49 to Z.

Routine provides for substitutions of other types as specified by sponsor, using substitution matrices up to size c by c.

9. COMBINED STREAMS.

ROUTINE. Forms combined streams.

PARAMETERS. Must specify the interval of spacing.

DESCRIPTION. Combines selected derived streams or sponsor-submitted input streams into one long stream, which will then be available for testing. The streams are combined in some prescribed order at some fixed interval of spacing, j ($100 \leq j \leq 2000$), gaps between streams being filled in with dits. For example, for interval 500 the first stream will occupy positions 0001 to L_1 (its length); the second, 0501 to (0500 plus L_2), etc. Each gap, of length (500-L_i) between the i-th substream and the succeeding one, will be filled in with dits.

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

REMARKS: Other routines may be added to the SUPERSTETH collection at any time.

Present plans call for the addition of

8.1.3 Sigma streams

4.5 Prime factors of offsets between polygraphic repeats.

4.6 Delta IC and sigmage of polygraphic repeats on various cuts.

For example, trinomic repeats will be counted on cuts 1, 2, and 3 separately as well as overall.

These separate polygraphic counts will be provided for polygraphs of length 3 to 6 for alphabets of size 10 or greater.

For alphabet size c= 2, provision may be made for separate n-bit counts up to size n= 20.

DECLASSIFIED
NW55645
AUTHORITY
BY LJ NARA DATE: 4/25

APPENDIX A

THE DELTA I.C.

(Explanation and examples of the use and interpretation of the Delta I.C. statistic and its sigmage-for monographs, digraphs, ..., polygraphs.)

DECLASSIFIED
NW5561649
AUTHORITY
BY LJ NARA DATE: 4/25

APPENDIX A

Delta I.C. (Index of Coincidence)

The delta I.C. statistic is used to measure the roughness of a polygraphic distribution (that is, of a frequency count of polygraphs of any size n where n = 1,2,3,... .) Monographs (n = 1), digraphs (n = 2), trigraphs (n = 3), tetragraphs (n = 4), pentagraphs (n = 5), etc. are all included here under the name polygraphs or n-graphs. An n-graphic frequency count is usually obtained by counting the n-graphs beginning at every position of the stream (overall), or by counting only the n-graphs which appear at an interval n.

EXAMPLE: For the stream

P H D Q Z J R I P E L S . . . ,

the overall frequency count of 4-graphs (tetragraphs) contains the 4-graphs PHDQ, HDQZ, DQZJ, QZJR, ZJRI, JRIP, etc.

The frequency count of 4-graphs at interval 4, beginning at position 1, contains the 4-graphs PHDQ, ZJRI, PELS, etc. There are three other 4-graphic frequency counts at interval 4 which begin at positions 2, 3, and 4, respectively. (The combined frequency count of the n different n-graphic counts at interval n is (of course) the same as the overall n-graphic count).

The digraphic (2-graphic) frequency count at interval 2, beginning at position 1, is called "on-cut"; while the count beginning at position 2 is called "off-cut". For the above stream, the on-cut count includes the digraphs PH, DQ, ZJ, ..., and the off-cut includes HD, QZ, JR, ...

For codes, the n-graphic frequency count usually desired is where n is equal to the group size of the code. An n-graphic frequency count might also be made for a stream resulting either from directional differences or minor differences of two messages, or for any streams derived from given messages.

The delta I.C. of an n-graphic distribution is the quotient of the actual (or observed) number of n-graphic hits (repeats or coincidences) divided by the expected number of hits. It always has the formula

$$(1) \qquad \delta_n = \frac{\sum_{i=1}^{K} \frac{f_i(f_i-1)}{2}}{\frac{N(N-1)}{2} \cdot \frac{1}{K}}$$

where δ_n = the n-graphic delta I.C.

N = the total number of n-graphs counted

K = the number of categories (that is, the total number of <u>different</u> n-graphs that are possible).

f_i = the frequency of the i-th category (or the i-th distinct n-graph).

(NOTE: For a given stream, the number N usually will not be the same for trigraphs as it is for digraphs, etc.)

DECLASSIFIED
NW556449
AUTHORITY
BY LJ NARA DATE: 4/25

SECRET

Equivalent formulas for the n-graphic delta I.C. are

$$(2)\qquad \delta_n = \frac{K \cdot \left(\sum_{i=1}^{K} f_i(f_i-1)\right)}{N(N-1)} \quad * \text{ and}$$

$$(3)\qquad \delta_n = \frac{K}{N(N-1)}\left[\left(\sum_{i=1}^{K} f_i^2\right) - N\right].$$

Each of these formulas will give the same value for the n-graphic delta I.C. The first formula, (1), will be used in the examples because it displays the actual number of observed hits and the number of expected hits. The numerator, $\sum_{i=1}^{K} \frac{f_i(f_i-1)}{2}$, is the number of observed hits. The denominator is the expected number of random hits, which is the total number of comparisons of n-graphs (in pairs) times the probability, at random, that any two n-graphs are identical.

For most applications, where it is possible for each character of the basic alphabet to appear in any of the n positions of an n-graph, the number K is equal to c^n, where c is the size of the alphabet.

EXAMPLES:

n = 2 (digraphs):

If c = 10 (digits), then the number of categories is $K = c^2 = 10^2 = 100$.

If c = 26 (26-letter alphabet), then $K = 26^2 = 676$.

n = 4 (tetragraphs):

If c = 10, then $K = 10^4 = 10{,}000$.

If c = 26, then $K = 26^4 = 456{,}976$.

n = 1 (monographs):

In these cases $K = c^1 = c$, the alphabet size.

There are situations where the number of categories, K, is not a power of the alphabet size, c. In such cases the number of categories must be determined separately. For example, for minor differences of a 4-digit code, the number of possible categories is K = 5008. As another example, if a 4-digit code contains only 6 different digits in the first position, then $K = 6 \times 10^3 = 6000$.

Having determined the correct number of categories, K, then the expected number of n-graphic repeats is given by $\frac{N(N-1)}{2} \cdot \frac{1}{K}$, the denominator of formula (1); and the delta I.C. for n-graphs can be calculated directly from the n-graphic frequency count by either of formulas (1), (2), or (3).

* The expression $\sum_{i=1}^{K} f_i(f_i-1)$ is often denoted by ϕ_o (phi observed), and the expression $N(N-1) \cdot \frac{1}{K}$ by ϕ_r (phi random). Then $\delta_n = \frac{\phi_o}{\phi_r}$.

SECRET

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

The SUPERSTETH program presently contains routines that calculate the delta I.C. for

monographs (Routines 1.2.1, 1.4.2, and 5.2) and digraphs (Routines 2.1.2 and 4.1).

Routine 4.1 calculates the delta I.C. for digraphs, trigraphs, ..., P-graphs; where P is the smallest integer such that the expected number of p-graphic repeats, $\frac{N(N-1)}{2} \cdot \frac{1}{c^P}$, is less than or equal to 20.

EXAMPLE I - Monographic Delta I.C. and Sigmage.

Given the following monographic frequency count:

(c = 26)

A	48	N	50
B	34	O	32
C	55	P	39
D	42	Q	34
E	51	R	34
F	29	S	44
G	37	T	50
H	36	U	49
I	37	V	39
J	39	W	38
K	42	X	44
L	39	Y	33
M	55	Z	56
		TOTAL	1086 = N

The number of categories K is the same as the alphabet size c = 26. The delta I.C. formula

(1) $$\delta = \frac{\sum_{i=1}^{K} \frac{f_i(f_i-1)}{2}}{\frac{N(N-1)}{2} \cdot \frac{1}{K}}$$ becomes

$$\delta = \frac{\sum_{i=1}^{26} \frac{f_i(f_i-1)}{2}}{\frac{N(N-1)}{2} \times \frac{1}{26}},$$ where

$f_1 = 48$, the frequency of A,

$f_2 = 34$, the frequency of B, ...

$f_{26} = 56$, the frequency of Z.

DECLASSIFIED
NW5556 45
AUTHORITY
BY LJ NARA DATE: 4/25

~~SECRET~~

The numerator of the delta I.C. formula, the observed or actual number of monographic repeats, is calculated as follows:

$$\sum_{i=1}^{26} \frac{f_i(f_i-1)}{2} = \frac{48\times47}{2} + \frac{34\times33}{2} + \frac{55\times54}{2} + \cdots$$

$$+ \frac{33\times32}{2} + \frac{56\times55}{2} = 22{,}895.$$

The denominator of the delta I.C. formula, the expected number of monographic repeats, is

$$\frac{N(N-1)}{2} \times \frac{1}{26} = \frac{1086\times1085}{2} \times \frac{1}{26} = 22{,}659.8$$

Then the monographic delta I.C. is

$$\delta = \frac{\sum_{i=1}^{26} \frac{f_i(f_i-1)}{2}}{\frac{N(N-1)}{2}\times\frac{1}{26}} = \frac{22{,}895}{22{,}659.8} = 1.0104.$$

The sigmage, the number of sigmas or standard deviations, is equal to the bulge (the difference between the calculated value of the delta I.C. and its expected value), divided by the value of the standard deviation. The standard deviation, or sigma, is given by the formula

$$\sigma = \frac{\sqrt{2(K-1)}}{\sqrt{N(N-1)}}.$$

If N is used as an approximation for $\sqrt{N(N-1)}$ (the approximation is good if N is large), the formula for sigma becomes

$$\sigma = \frac{\sqrt{2(K-1)}}{N}.$$

The sigmage of the delta I.C. has the formula

$$S(\delta) = \frac{\text{bulge}}{\sigma} = \frac{\text{bulge}}{\frac{\sqrt{2(K-1)}}{N}}$$

For the present example, where K = c = 26, N = 1086, δ = 1.0104 and the random expected value of δ is 1.0, the sigmage of the monographic delta I.C. is

$$S(\delta) = \frac{1.0104 - 1.0}{\frac{\sqrt{2(26-1)}}{1086}} = \frac{0.0104}{\frac{\sqrt{50}}{1086}} = \frac{0.0104}{\frac{7.071}{1086}}$$

$$= \frac{0.0104}{0.00651} = 1.60 \text{ sigmas.}$$

~~SECRET~~

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

SECRET

The sigmage may be interpreted in terms of the probability, at random, of obtaining a delta I.C. that has a sigmage at least as large as the calculated value, 1.60. This probability may be obtained from a table of probabilities for the I.C. Such a table is printed on page 65, and a short extract for c = 26, is presented here.

TABLE I

The probability, P, of obtaining a sigmage greater than or equal to S(δ) is P.

(c = 26)

S(δ)	P
0.0	0.463
0.2	0.386
0.4	0.316
0.6	0.254
0.8	0.201
1.0	0.156
1.2	0.120
1.4	0.0902
1.6	0.0671
1.8	0.0493
2.0	0.0358
2.2	0.0256
2.4	0.0182
2.6	0.0127
2.8	0.00885
3.0	0.00609
3.2	0.00415
3.4	0.00280
3.6	0.00188
3.8	0.00125
4.0	0.00083
4.2	0.00054
4.4	0.00035
4.6	0.00023
4.8	0.00015
5.0	0.00009
5.2	0.00006
5.4	0.00004
5.6	0.00002
5.8	0.000015
6.0	0.000009
6.2	0.000006

A sigmage of 1.6 has the associated probability 0.0671. This means that one may expect to find a sigmage (for the delta I.C., when c = 26) as large as 1.6, about 67 times out of every 1000 trials (or about one time out of every 15 trials) - not a very significant result.

SECRET

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

(Equivalent probabilities may be obtained by calculating chi-squared from the formula $\chi^2 = (N-1)(\delta-1)+(c-1)$, and finding the probability associated with this value of χ^2 in a chi-squared table with $(c-1)$ degrees of freedom. For this example, $N=1086$, $(\delta-1)=0.0104$ and $(c-1)=25$ and so $\chi^2=(1085 \times 0.0104)+25=36.29$, with 25 degrees of freedom.)

EXAMPLE II Digraphic Delta I.C. and Sigmage.

Given the following on-cut dinomic frequency count:

	1	2	3	4	5	6	7	8	9	0
1	6	3	2	0	4	3	0	3	10	5
2	3	0	5	7	3	1	6	2	6	4
3	3	4	5	4	5	6	9	5	7	0
4	0	5	3	3	1	4	11	5	7	1
5	1	8	4	5	3	0	10	6	12	3
6	4	3	4	4	3	2	5	7	12	7
7	0	2	15	13	8	6	11	3	16	8
8	4	3	3	0	5	3	12	7	13	7
9	13	2	6	5	10	16	17	18	8	6
0	4	3	3	4	9	6	4	1	4	5

TOTAL $M=547$

(NOTE: In this paper the letter M is used instead of N for the total number of digraphs.)

The number of categories K is the square of the "alphabet" size $c=10$; that is, $K=c^2=10^2=100$. The delta I.C. formula

(1) $$\delta = \frac{\sum_{i=1}^{K} \frac{f_i(f_i-1)}{2}}{\frac{N(N-1)}{2} \cdot \frac{1}{K}}$$ becomes (for digraphs)

$$\delta_2 = \frac{\sum_{i=1}^{100} \frac{f_i(f_i-1)}{2}}{\frac{M(M-1)}{2} \times \frac{1}{100}}$$ where

f_i = the frequency of the i-th digraph, and M = the total number of digraphs counted.

(NOTE: The summation $\sum_{i=1}^{100} \frac{f_i(f_i-1)}{2}$ means that all the terms $\frac{f_i(f_i-1)}{2}$ are added, one term for each of the 100 different digraphs.)

The numerator of the delta I.C. formula, the observed or actual number of digraphic repeats, is calculated as follows:

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

SECRET

$$\sum_{i=1}^{100} \frac{f_i(f_i-1)}{2} = \frac{6X5}{2} + \frac{3X2}{2} + \frac{2X1}{2} + \emptyset + \frac{4X3}{2} + \frac{3X2}{2}$$
$$+ \emptyset + \frac{3X2}{2} + \frac{10X9}{2} + \frac{5X4}{2} + \frac{3X2}{2} + \emptyset + \frac{5X4}{2} +$$
$$\cdots + \emptyset + \frac{4X3}{2} + \frac{5X4}{2} = \frac{4080}{2} = 2040.$$

The denominator of the delta I.C. Formula, the expected number of random digraphic repeats, is

$$\frac{M(M-1)}{2} \times \frac{1}{100} = \frac{547X546}{2} \times \frac{1}{100} = 1493.31.$$

Then the digraphic delta I.C. is

$$\delta_2 = \frac{\sum_{i=1}^{100} \frac{f_i(f_i-1)}{2}}{\frac{M(M-1)}{2} \times \frac{1}{100}} = \frac{2040}{1493.31} = 1.3661.$$

The standard deviation, σ, is given by the formula

$$\sigma = \frac{\sqrt{2(K-1)}}{M} = \frac{\sqrt{2(100-1)}}{547} = \frac{14.071}{547} = 0.02572,$$

where M is an approximation for $\sqrt{M(M-1)}$.

The sigmage of the delta I.C. is given by the formula

$$S(\delta) = \frac{bulge}{\sigma} = \frac{bulge}{\frac{\sqrt{2(K-1)}}{M}} = \frac{0.3661}{0.02572} = 14.23$$

sigmas.

The sigmage may be interpreted in terms of the probability, at random, of obtaining a delta I.C. that has a sigmage at least as large as the calculated value, 14.23. This probability may be obtained (with a sufficient degree of accuracy* for most applications) from a probability table for the normal distribution. Such a table is printed on page 66 ; but the probability for a sigmage of 14.2 is not even shown. This means that the probability, at random, of finding a digraphic distribution this rough is much less than .0000001 (that is, less than once out of every 10 million trials) - a highly significant result!

(Note: In SUPERSTETH routine 2.1.2.1, the standard deviation is changed to $\frac{\sqrt{2c(c+2)}}{M}$ for overall digraphic counts in order to remove much of the effect of the dependence of overlapping digraphs.)

* Footnote: Greater accuracy can be obtained by the use of tables of the Incomplete Gamma Function.

SECRET

DECLASSIFIED
NW5561/49
AUTHORITY
BY LJ NARA DATE: 4/25

EXAMPLE III

Polygraphic Delta I.C. and Sigmage and The Index of Cohesion.

Given a stream of 503 digits (no dits), with the following n-graphic (overall) delta I.C.'s for n=1 (mononomes), n=2 (dinomes), n=3 (trinomes) and n=4 (tetranomes). The n-graphic delta I.C.'s are computed directly from the n-graphic frequency count by means of the formula:

$$\delta_n = \frac{\text{number of observed n-graphic repeats}}{\text{number of expected random n-graphic repeats}}$$

$$= \frac{\sum_{i=1}^{K} \frac{f_i(f_i-1)}{2}}{\frac{N(N-1)}{2} \times \frac{1}{K}} \quad \text{where}$$

the number of categories $K = c^n = 10^n$, and the total number, N, of n-graphs counted are as follows:

$K = 10^1 = 10$, $N = 503$ for $n = 1$;

$K = 10^2 = 100$, $N = 502$ for $n = 2$;

$K = 10^3 = 1000$, $N = 501$ for $n = 3$;

and $K = 10^4 = 10{,}000$, $N = 500$ for $n = 4$.

(f_i is the frequency of the i-th category).

The n-graphic delta I.C.'s were found to be (for mononomes, dinomes, trinomes, and tetranomes respectively):

$$\delta_1 = 1.01$$

$$\delta_2 = 1.08$$

$$\delta_3 = 1.09$$

$$\delta_4 = 1.84 \ .$$

The sigmage of the n-graphic delta I.C. is equal to the "bulge" (the difference between the observed and expected values of the I.C.) divided by the standard deviation (sigma). The SUPERSTETH program uses $\frac{\sqrt{2(K-1)}}{N}$ as an approximation for sigma.

The formula for the sigmage of the n-graphic delta I.C. is

$$S(\delta_n) = \frac{\delta_n - (\delta_{n-1} \times \delta_1)}{\frac{\sqrt{2(K-1)}}{N}} \quad \text{where}$$

δ_n is the delta I.C. for n-graphs,

δ_{n-1} is the delta I.C. for (n-1)-graphs,

and δ_1 is the delta I.C. for monographs.

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

(Note that the expected value of δ_2 is taken as $\delta_1 \times \delta_1 = (\delta_1)^2$, the expected value of δ_3 is $\delta_2 \times \delta_1$, the expected value of δ_4 is $\delta_3 \times \delta_1$, etc. For δ_1 the random expected value, 1.0, is used).

The expression $\sqrt{2(K-1)}$ is, for the respective n-graphs, as follows:

$\sqrt{2(K-1)} = \sqrt{18} = 4.2426$ for n=1;

$\sqrt{198} = 14.071$ for n=2;

$\sqrt{1998} = 44.70$ for n=3;

and $\sqrt{19,998} = 141.4$ for n=4.

The respective sigmages are computed as follows:

$$S(\delta_1) = \frac{1.01-1.0}{\frac{4.2426}{503}} = \frac{0.01}{.008435} = 1.19 \text{ sigmas,}$$

$$S(\delta_2) = \frac{1.08-1.0201}{\frac{14.071}{502}} = \frac{0.0599}{.02803} = 2.14 \text{ sigmas,}$$

$$S(\delta_3) = \frac{1.09-1.0908}{\frac{44.7}{501}} = \frac{-0.0008}{.08922} = -0.01 \text{ sigmas,}$$

$$S(\delta_4) = \frac{1.84-1.1009}{\frac{141.4}{500}} = \frac{0.7391}{0.2828} = 2.61 \text{ sigmas.}$$

The sigmages may be interpreted in terms of the probability of obtaining delta I.C.'s which have sigmages at least as large as the values calculated. For monographs it is recommended that a probability table of the I.C. distribution, such as the one on page 65, be used. For digraphs, trigraphs, tetragraphs, etc. a table of probabilities of the normal distribution (such as the one on page 69) may be used. (The approximation, to the true probability, given by the normal table is sufficiently accurate for most cryptanalytic purposes. The accuracy increases as the number of categories, K, increases).

Using the I.C. table (page 65) to interpret the sigmage of the monographic delta I.C., $S(\delta_1)$, we find in the column for c=10 that a sigmage of 1.2 has the associated probability 0.119 or about 0.12. This means that about 12 messages out of every 100 will have a mononomic sigmage as large as 1.2 (at random) - an insignificant result.

DECLASSIFIED
NW5564S
AUTHORITY
BY LJ NARA DATE: 4/25

Using the normal table (page 66) to interpret the sigmage of the digraphic I.C., $S(\delta_2)$ we find that a sigmage of 2.1 has the associated probability 0.0179 or about 0.018. This means that about 18 messages out of every 1000 will have a digraphic sigmage as large as 2.1 (at random) a more interesting result.

The value of $S(\delta_3)$ is negative (indicating that the value of δ_3 is less than its expected value), but is so close to zero as to be very insignificant.

Using the normal table to interpret the sigmage of the tetragraphic I.C., $S(\delta_4)$, we find that a sigmage of 2.6 has the probability 0.0047 or about 0.005. This means that only about 5 messages out every 1000 at random will have a tetragraphic sigmage as large as 2.6 - a quite significant result (unless we had to test many messages in order to find one like this)!

One might conclude from these results that the system is tetranomic or possibly dinomic. A more definite conclusion might be reached in the following way. The number of tetranomic repeats is *23, where $\frac{500 \times 499}{2} \times \frac{1}{10^4} = 12.5$ are expected (at random).

(NOTE that $\delta_4 = \frac{23}{12.5} = 1.84$). Now when 12.5 hits are expected one may usually expect the number of random hits to range somewhere between about 6 to about 19. Thus some number of hits, ranging between 17 and 4 should be causal. The causal hits may often be determined by finding a consistent interval at which they appear or by finding a common factor of the offset distance between hits.

Another means of determining the most significant n-graph is provided by the index of cohesion. The index of cohesion for n-graphs is the ratio of the calculated or observed value of the n-graphic delta I.C. to its expected value.

For this example the index of cohesion for each of the various n-graphs is:

$$i_1 = \frac{\delta_1}{1} = 1.01$$

$$i_2 = \frac{\delta_2}{\delta_1 \times \delta_1} = \frac{1.08}{1.0201} = 1.0587$$

$$i_3 = \frac{\delta_3}{\delta_2 \times \delta_1} = \frac{1.09}{1.0908} = 0.9993$$

$$i_4 = \frac{\delta_4}{\delta_3 \times \delta_1} = \frac{1.84}{1.1007} = 1.6714$$

Note that the index of cohesion for tetranomes is outstanding.

FOOTNOTE: * The number of n-graphic repeats can be easily calculated from the SUPERSTETH-printed value of δ_n, by using the following formula:

$$\#\text{ of n-graphic repeats} = \left(\frac{N(N-1)}{2} \times \frac{1}{c^n}\right) \times \delta_n .$$

In this example, the number of tetranomic repeats is

$$(12.5) \times 1.84 = 23.$$

DECLASSIFIED
NW5516 49
AUTHORITY
BY LJ NARA DATE: 4/25

To be assured that the apparent significance of tetranomes does not come from an excess of pentanomes, 6-nomes, etc. one should check the statistics for longer repeats. For this example, the SUPERSTETH program printed 19 tetranome repeats and 2 pentanome repeats but no longer ones. (The pentanome repeats each contain 2 tetranome repeats. Thus there are 19+(2×2) tetranome repeats altogether).

The pentanomic statistics may be computed as follows:

$$\delta_5 = \frac{2}{\frac{N(N-1)}{2} \times \frac{1}{105}} = \frac{2}{1.242} = 1.6103,$$

$$s(\delta_5) = \frac{1.6103 - 1.8584}{\frac{447.21}{499}} = \frac{-0.2481}{0.8962} = -0.28 \text{ sigmas},$$

$$i_5 = \frac{1.6103}{1.84 \times 1.01} = \frac{1.6103}{1.8584} = 0.866$$

Thus the pentanomic repeats are very insignificant. Since there are no 6-nomic repeats, the delta I.C. for 6-nomes, 7-nomes, ..., are all zero.

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

APPENDIX B

THE CHI-SQUARED STATISTIC

Written by Edwin J. Hughes

DECLASSIFIED
NW5564S
AUTHORITY
BY LJ NARA DATE: 4/25

SECRET

THE CHI-SQUARED STATISTIC

OUTLINE

1. THE FUNCTION OF DIAGNOSIS.
   (What diagnosis is supposed to do.)

2. THE DIAGNOSTIC PROCEDURE.
   (How diagnosis is effected.)

3. DIAGNOSIS USING CHI-SQUARED.
   (What the Chi-squared test does.)

4. CHI-SQUARED AS A MEASURE OF FIT.
   (How Chi-squared does what is required to measure fit.)

5. INTERPRETATION OF THE CHI-SQUARED SCORE.
   (How Chi-squared scores run: (a.) The larger the Chi-squared the worse the fit, with a fixed number of categories. (b.) The greater the number of categories the larger the Chi-squared, for the same degree of fit.)

6. DEGREES OF FREEDOM OF THE CHI-SQUARED STATISTIC.
   (Why degrees of freedom for Chi-squared statistics.)

7. THE $\sqrt{2\chi^2}$ STATISTIC.
   (How to relate χ^2 to probabilities via $\sqrt{2\chi^2}$.)

8. CAUTIONS.
   (Required for the Chi-squared test to be valid.)

9. EXAMPLES.

SECRET

DECLASSIFIED
NW55645
AUTHORITY
BY LJ NARA DATE: 4/25

THE FUNCTION OF DIAGNOSIS.

An important problem in Cryptanalysis is to find the method of generation of a stream of cipher text or key. The plan of attack is first to find something that requires explaining and then to develop an explanation. It is a problem of diagnosis to find something that requires explaining: a property of the data that is non-random, or causal, presumably arising from a cause related to the generation of the stream.

THE DIAGNOSTIC PROCEDURE.

Diagnosis is often effected in the following way. A statistic is evaluated to measure some property of the data. The hypothesis is made that, with respect to the property being measured, the data is random; that is, that the property is not a causal effect, and there is nothing to be explained. The measurement taken is compared with the range for such measurements on random data. Should this measurement be within the range for random, so that as a measurement on random data it is not unlikely, then the hypothesis cannot be rejected and no decision can be made as to whether or not the property under investigation arises from some cause. However, should this measurement be well beyond the range for random, so that as a measurement on random data it is extremely unlikely, then the hypothesis is rejected, and it is assumed that the measurement is non-random and that the property of the data under investigation arises from some cause.*

In this case there has been found something that requires explaining.

Note that the argument used in this second case is very simple:
If there is no cause, then the result of the test will be within the range of random results.
But the result of the test is not within the range of random results.
Therefore it is not true that there is no cause. (Hence causal.)

DIAGNOSIS USING CHI-SQUARED.

A stream of cipher text or key is considered a sample of data and can be characterized by various frequency counts or distributions involving the characters in the stream: monographic, digraphic, trigraphic distributions; the density distribution; distributions of dots and crosses in the levels; and other distributions. The Chi-squared statistic is used to measure the fit of a distribution observed in the sample of data to a distribution expected from a random hypothesis. The statistic is evaluated, that is, the measurement is taken, by summing category by category deviations of one distribution from the other. Thus a Chi-squared score is a number which is a sum of deviations. By the size of the Chi-squared score, one can judge how well or how poorly the two distributions match. A large value of Chi-squared arises from large deviations and indicates a poor match or fit; a small value arises from small deviations and indicates a good fit.

* It should be observed that measurements which are too near to expected random may also arise from a cause: a deliberate effort to make the data perfectly random. (e.g. a monographic distribution very close to uniformly flat).

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

It is important to understand that the method of attack in using the Chi-squared statistic to find non-random or causal properties involves the comparison of the observed Chi-squared score with the range of random Chi-squared scores: the comparison of the Chi-squared score for the fit of the observed distribution to the expected random distribution* with the range of Chi-squared scores for the fit of a random distribution to the expected random distribution.

With an appreciation of some advanced topics of Mathematical Statistics, one would know that the range of Chi-squared scores for the fit of random to expected random is that presented in a table of the Chi-squared distribution.** Such a table relates Chi-squared scores to probabilities, giving for various Chi-squared scores the probability that the Chi-squared statistic will assume a value that large or larger. Thus, by reference to the table, one can find whether or not the Chi-squared score for the fit of the distribution observed to expected random is within the range of Chi-squared scores for the fit of random to expected random. From this, one can judge whether or not the random hypothesis can be rejected.

CHI-SQUARED AS A MEASURE OF FIT.

Given are two frequency distributions: an observed distribution and an expected distribution. The expected distribution consists of the same total count as the observed, and is distributed into the same number of categories, but distributed in accordance with a random hypothesis, which is to be tested. It is desired to form some measure of the fit of the observed distribution to the expected distribution.

It is intuitively clear that a measure of fit (perhaps, lack of fit) should involve a sum of deviations. Thus a measure of the fit of the observed to the expected distribution would be the sum, over all categories, of the deviation in the category, i.e. observed minus expected count.

$(o_i - e_i)$ = deviation in category i

$\sum_{i=1}^{c} (o_i - e_i)$ = sum, over all categories, of the deviations.

A bit of thought suggests that a negative deviation should not be cancelled against, but rather added to a positive deviation; and therefore a more suitable measure of fit would be the sum, over all categories, of the absolute value of the deviation in the category.

$\sum_{i=1}^{c} |(o_i - e_i)|$ = sum, over all categories, of the absolute value of the deviations.

Further thought suggest that a deviation of 2 when 10 is expected ought to be considered more serious than a deviation of 2 when 100 is expected. In terms of probabilities,

P (deviation 2, expect 100) > P (deviation 2, expect 10).

* The term "expected distribution" is not ordinarily used. Here "expected random distribution" is used to mean the distribution which has, in each category, that count expected from a random hypothesis.

** Contrast "distribution" as meaning a list of categories with corresponding frequency counts, e.g. digraphic distribution, with "distribution" as meaning a list of scores with corresponding probabilities, e.g. Chi-squared distribution.

DECLASSIFIED
NW55649
AUTHORITY
BY NARA DATE: 4/25

SECRET

Thus a better measure of fit is suggested: the sum, over all categories, of the absolute value of the deviation per unit of expected value in the category, something of a sum of percentage deviations.

$$\sum_{i=1}^{c} \frac{|(o_i - e_i)|}{e_i} = \text{sum, over all categories, of the absolute value of the deviation per unit of expected value.}$$

Still further thought reveals that a deviation of 20 when 100 is expected is more serious than a deviation of 2 when 10 is expected. In terms of probabilities.

$$P \text{ (deviation 20, expect 100)} < P \text{ (deviation 2, expect 10).}$$

Thus, equal percentage deviations are not equally serious. An intuitive approach to an understanding of this is to consider a deviation of 20 when 100 is expected as a ten-fold occurrence of a deviation of 2 when 10 is expected. Clearly, a multiple occurrence of an unlikely event is even more unlikely.

Just as there is an expected value, e_i, in a category, there is also an expected deviation. It is called the standard deviation and is denoted by σ_i. The deviation per unit of expected deviation is called the sigmage. Whereas equal percentage deviations are not equally likely, equal sigmage deviations are. Hence, a better measure of fit would be the sum, over all categories, of the absolute value of the sigmage, that is, the absolute value of the deviation per unit of expected deviation.

$$\sum_{i=1}^{c} \frac{|(o_i - e_i)|}{\sigma_i} = \sum_{i=1}^{c} |(\text{sigmage})_i|$$

$$= \text{sum, over all categories, of the absolute value of the deviation per unit of expected deviation.}$$

In order for a statistic to be useful in diagnostic testing there must be a way to relate values of the statistic to probabilities. That is, the distribution of the statistic must be known. Sigmages are approximately distributed according to the Normal distribution, but a sum of absolute values of sigmages is not distributed according to the Normal, or any known distribution. However, the sum of the squares of sigmages is distributed according to a known distribution, the Chi-squared distribution. A statistic which is the sum of the squares of sigmages satisfies both requirements of providing a measure of fit and having a known distribution. This is the Chi-squared statistic.

$$\chi^2 = \sum_{i=1}^{c} (\text{sigmage}_i)^2 = \sum_{i=1}^{c} \left[\frac{(o_i - e_i)}{\sigma_i}\right]^2$$

SECRET

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

SECRET

The standard deviation of a category count is approximately equal to the square root of the expected category count, in accordance with the formulas given below without explanation.

$$\sigma_i = \sqrt{N p_i (1-p_i)} \doteq \sqrt{N p_i} = \sqrt{e_i} \quad \text{if } p_i \text{ is small.}$$

Hence the final expression for the measure of fit, Chi-squared, is

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i} .$$

INTERPRETATION OF THE CHI-SQUARED SCORE.

In order to understand the use of the Chi-squared statistic one must have some appreciation of the range of values it can assume.

1. With a fixed number of categories, the larger the Chi-squared score, the worse the fit.

Fundamentally the Chi-squared statistic is a sum of deviations, and therefore, with a fixed number of categories, the larger the value of Chi-squared, the larger the deviations, and the worse the fit.

Thus, in the example below, observed distribution A is a better fit to expected distribution E than observed distribution B.

A	9	9	10	11	11
E	10	10	10	10	10

$$\chi^2_A = \frac{4}{10} = 0.4$$

B	8	9	10	11	12
E	10	10	10	10	10

$$\chi^2_B = \frac{10}{10} = 1.0.$$

2. The greater the number of categories, the larger the Chi-squared score for the same degree of fit.

Consider a Chi-squared statistic for measuring the fit of a distribution of 2 categories to expected random and a Chi-squared statistic for measuring the fit of a distribution of 10 categories to expected random. Let both observed distributions have 100 counts. The expected random (uniformly flat) distributions are respectively 50 - 50 and 10, 10, 10, 10, 10 - 10, 10, 10, 10, 10. One can consider the expected distribution for the 10-category distribution as specifying a refinement on the expected distribution for a 2-category distribution, such that each of the 2 categories is divided into 5 sub-categories into which the 50 counts must be distributed equally, rather than in an arbitrary manner. From these and similar considerations it should be clear that the greater the number of categories, the greater are the demands for a good fit. Thus the greater the number of categories, the larger will be the Chi-squared score for the same degree of fit. In the example below, observed distribution B has approximately the same degree of fit to expected random as observed distribution A, although the Chi-squared score for B, with 10 categories, is 16.8, while that for A, with 2 categories, is 4.0.

SECRET

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

SECRET

A	40	60
E	50	50

$\chi^2_A = 4.0$

B	7	13	7	13	6	14	5	15	5	15
E	10	10	10	10	10	10	10	10	10	10

$\chi^2_B = 16.8$

From tables of the Chi-squared distribution (Table 1, below) it is found that the probability that a Chi-squared score be as large or larger than 4.0 when there are 2 categories is approximately the same as the probability that a Chi-squared score be as large or larger than 16.8 when there are 10 categories. These cases are equally likely and occur about 5 percent of the time. Thus only 5 percent of the times that a random distribution is compared with the uniformly flat distribution will there be as bad or worse a fit as was observed in the examples above.

DEGREES OF FREEDOM OF THE CHI-SQUARED STATISTIC.

The Chi-squared statistic is a number which measures the fit of an observed distribution to an expected distribution arising from a random hypothesis. A comparison is made of the Chi-squared score for the fit of what is observed to expected random with the range of Chi-squared scores for the fit of random to expected random. This comparison is expressed in terms of probabilities by reference to tables of the Chi-squared distribution. The tables give for various scores the probability that the Chi-squared statistic will assume a value that large or larger.

The range of Chi-squared scores for the fit of random to expected random is dependent upon the number of categories in the distribution. That is to say, there is a different Chi-squared distribution for each number of categories. This invites an investigation of the relationship between the range of Chi-squared scores and the number of categories in the distribution.

One might compare the fit of an observed c-character monographic distribution with the distribution which has, in each category, that count expected by a random selection from a flat distribution of the same total count. In trying to fit a random distribution to this type of expected random distribution, restrictions are imposed which allow only (c-1) degrees of freedom in specifying the distribution of the total count into (c-1) of the c categories, the last count being determined by the (c-1) others and the observed total count.

One might compare the fit of an observed c^2 - character digraphic distribution with the distribution which has, in each category, that count expected by a random selection of digraphs from the two observed monographic distributions in the first and second positions (row and column sums of the c x c digraphic distribution matrix). In trying to fit a random distribution to this type of expected random distribution, restrictions are imposed which allow only $(c-1)^2$ degrees of freedom in specifying the distribution of the total count into $(c-1)^2$ of the c^2 categories, the remaining (2c-1) counts being determined by the (c-1) others and the observed row and column sums. (Thus in the c x c matrix the right-hand column and the bottom row, (2c-1) counts are determined by the (c-1) counts in the remainder of the matrix and the row and column sums.)

SECRET

DECLASSIFIED
NW55645
AUTHORITY
BY LJ NARA DATE: 4/25

SECRET

From these considerations, the first statistic is called a Chi-squared statistic with (c-1) degrees of freedom and the second a Chi-squared statistic with $(c-1)^2$ degrees of freedom. Each Chi-squared statistic is associated with a certain number of degrees of freedom, which is determined by the method of formation of the expected distribution (and which is often difficult to determine).

Each Chi-squared statistic is distributed according to a Chi-squared distribution with its number of degrees of freedom. Thus, what is true without qualification is that there is a distinct Chi-squared distribution for each number of degrees of freedom. That there is a Chi-squared distribution for each number of categories is true only in the sense that there is a distribution for each number of degrees of freedom. Below are given tables of the Chi-squared distribution for 1, 9, 25, and 31 degrees of freedom (which would be useful for Chi-squared tests on 2, 10, 26, and 32 category monographic distributions).

TABLE I

CHI-SQUARED DISTRIBUTION

The Chi-squared statistic, where there are ν degrees of freedom, assumes a value of X^2 or larger with a probability of p.

$\nu = 1$ X^2	$\nu = 9$ X^2	$\nu = 25$ X^2	$\nu = 31$ X^2	p
0.0000	0.66	6.7	9.8	0.9999
0.0000	1.2	8.7	12.2	0.999
0.0002	2.1	11.5	15.7	0.99
0.004	3.3	14.6	19.3	0.95
0.02	4.2	16.5	21.5	0.90
0.46	8.3	24.3	30.	0.50
2.7	14.7	34.	42.	0.10
3.8	16.9	38.	45.	0.05
6.6	21.7	44.	52.	0.01
10.8	27.9	53.	61.	0.001
15.1	34.	61.	69.	0.0001

A Chi-squared statistic can be characterized by its average or expected value. For any Chi-squared statistic the expected value is equal to the number of degrees of freedom. This is proved in the theory of Mathematical Statistics.

$$E(\chi^2) = \nu .$$

SECRET

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

SECRET

THE STATISTIC $\sqrt{2X^2}$

How to relate Chi-squared scores to probabilities by way of the $\sqrt{2X^2}$ statistic.

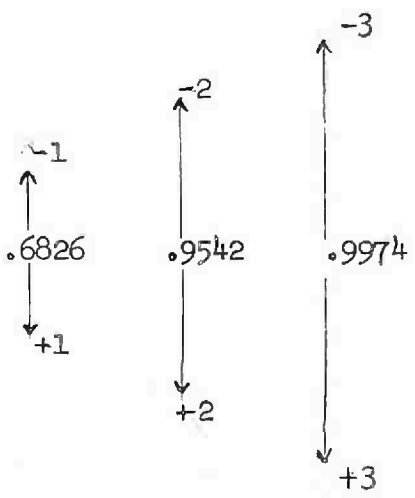
One can determine the likelihood, under some hypothesis, of the value assumed by any statistic by finding how far the value assumed * stands from its expected value. This can be expressed in terms of its sigmage: how many standard deviations or sigmas the statistic stands from its expected value. Probabilities can be assigned to the various possibilities that a statistic assume such and such a value amounting to so many standard deviations from its expected value. The range of values a statistic can assume, with the associated probabilities, is called its distribution. For those statistics which are distributed according to the Normal distribution, the probabilities corresponding to deviations are such that 68 percent of the time the statistic will assume a value within ±1 standard deviation of its expected value; 95 percent of the time within ±2 standard deviations; and 99.7 percent of the time within ±3 standard deviations.

TABLE 2

NORMAL DISTRIBUTION

A statistic distributed according to the Normal distribution will assume a value corresponding to S standard deviations or greater with a probability p.

S	p
-4.0	>0.9999
-3.0	0.9987
-2.5	0.9938
-2.0	0.9772
-1.5	0.9332
-1.0	0.8413
-0.5	0.6915
0.0	0.5000
+0.5	0.3085
+1.0	0.1587
+1.5	0.0668
+2.0	0.0228
+2.5	0.0062
+3.0	0.0013
+4.0	<0.0001



* The words "value assumed" or "assumed value" mean the same as "observed value".

SECRET

DECLASSIFIED
NW5564S
AUTHORITY
BY LJ NARA DATE: 4/25

SECRET

The Chi-squared statistic itself is not distributed according to the Normal distribution. However, if the number of degrees of freedom, ν , is 30 or greater, then a new statistic, $\sqrt{2X^2}$, is distributed approximately according to the Normal distribution.

For the statistic $\sqrt{2X^2}$:

Average or expected value: $E\left(\sqrt{2X^2}\right)=\sqrt{2\nu-1}$

Standard (average) deviation: $\sigma\left(\sqrt{2X^2}\right)=1$

Sigmage: $S\left(\sqrt{2X^2}\right)=\sqrt{2X^2}-\sqrt{2\nu-1}$.

Example. To relate Chi-squared scores to probabilities by way of the $\sqrt{2X^2}$ statistic.

The Chi-squared statistic for the fit of an observed 32-category distribution of 1000 characters to that uniformly flat distribution which has in each category $\frac{1}{32}(1000)$ counts assumes a value of 45. How unusual is this score? The range of values for the fit of a random 32-category distribution of 1000 characters to that uniformly flat (expected random) distribution is given in the table of the Chi-squared distribution with $(32-1)=31$ degrees of freedom. On comparison of the Chi-squared score of 45 with this distribution, one finds (Table 1)

$$P\ (X^2 \geq 45,\ 31 \text{ degrees of freedom}) \leq 0.05.$$

Here $E(X^2)=31$.

Interpretation: Only 5 percent of the Chi-squared scores for random vs. expected random are as large or larger than this Chi-squared score for observed vs. expected random.

Using instead the $\sqrt{2X^2}$ statistic, we have

$$S\left(\sqrt{2X^2}\right)=\sqrt{2X^2}-\sqrt{2\nu-1}$$

$$=\sqrt{2(45)}-\sqrt{2(32-1)-1}$$

$$S\left(\sqrt{2X^2}\right)=1.68.$$

The $\sqrt{2X^2}$ statistic will assume a value corresponding to 1.68 standard deviations or greater with a probability (Table 2)

$$0.0228 \leq P\left(S\left(\sqrt{2X^2}\right) \geq 1.68\right) \leq 0.0668.$$

From more extensive tables

$$P\left(S\left(\sqrt{2X^2}\right) \geq 1.68\right) \doteq 0.0465.$$

The interpreations is the same as above. Only 4.65 percent of the $\sqrt{2X^2}$ scores for random vs. expected random are as far from the expected $\sqrt{2X^2}$ score for random vs. expected random as this $\sqrt{2X^2}$ score for observed vs. expected random.

SECRET

DECLASSIFIED
NW55645
AUTHORITY
BY LJ NARA DATE: 4/25

CAUTIONS.

1. For valid results the Chi-squared statistic should not be used, without appropriate modification, for any test where the expected value in any category is less than 5. Those categories for which the expected value is less than 5 should be combined with other categories, one degree of freedom being subtracted for each category removed.

2. Chi-squared statistics for measuring the fit of an observed digraphic distribution to that expected from the observed monographic distributions in the first and second positions can involve terms of the form 0/0 if in the monographic distributions a category count is zero. For valid results, such terms 0/0 should be replaced by unity in the evaluation of Chi-squared.

EXAMPLE 1.

PROBLEM. To distinguish transposition cipher from random cipher by evaluating and interpreting the Chi-squared statistic for the fit of the monographic distribution to an expected random (uniformly flat) distribution.

OBSERVED CIPHER DISTRIBUTION.

N characters, c categories, with counts

$f_1, f_2, \ldots, f_c$ $\qquad f_1 + f_2 + \ldots + f_c = \sum_{i=1}^{c} f_i = N.$

EXPECTED UNIFORMLY FLAT DISTRIBUTION.

N characters, c categories, with counts

$e_1, e_2, \ldots, e_c$ $\qquad e_1 + e_2 + \ldots + e_c = \sum_{i=1}^{c} e_i = N.$

$e_1 = e_2 = \ldots = e_c = \frac{1}{c} \cdot N.$

EVALUATION OF CHI-SQUARED.

$$\chi^2 = \frac{(f_1 - e_1)^2}{e_1} + \frac{(f_2 - e_2)^2}{e_2} + \ldots + \frac{(f_c - e_c)^2}{e_c}$$

$$\chi^2 = \sum_{i=1}^{c} \frac{(f_i - e_i)^2}{e_i}.$$

$$\chi^2 = \sum_{i=1}^{c} \frac{\left(f_i - \frac{N}{c}\right)^2}{\frac{N}{c}} = \frac{c}{N}\left(\sum_{i=1}^{c} f_i^2\right) - N.$$

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

SECRET

NUMERICAL EXAMPLE.

Frequency distributions of two messages, with expected distributions.

First line is Message T.
Second line is Expected Message.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	TOTAL
18	4	8	11	30	7	4	10	18	2	2	9	6	21	19	7	1	20	15	22	8	7	4	1	5	1	260
10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	260

First line is Message R.
Second line is Expected Message.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	TOTAL
7	15	9	8	7	13	12	9	7	2	14	9	5	16	6	7	11	12	13	8	9	14	13	16	8	10	260
10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	260

$$\chi^2_T = \frac{(18-10)^2}{10} + \frac{(4-10)^2}{10} + \cdots + \frac{(1-10)^2}{10}$$

$$\chi^2_T = 155.6.$$

$$\chi^2_R = \frac{(7-10)^2}{10} + \frac{(15-10)^2}{10} + \cdots + \frac{(10-10)^2}{10}$$

$$\chi^2_R = 32.2.$$

INTERPRETATION. On reference to a table of the Chi-squared distribution with 25 = (26-1) degrees of freedom (Table 1).

$$P(X^2 \geq 155.6, 25 \text{ d.f.}) < 0.0001$$

$$0.10 < P(X^2 \geq 32.2, 25 \text{ d.f.}) < 0.50$$

$$P(X^2 \geq 32.2, 25 \text{ d.f.}) \doteq 0.16, \text{ from more extensive tables.}$$

Note that here the expected value of Chi-squared is equal to the number of degrees of freedom.

$$E(X^2) = 25.$$

SECRET

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

SECRET

The interpretation of the Chi-squared score 155.6 for Message T is as follows. Under the hypothesis that, with respect to the fit of its monographic distribution to a uniformly flat distribution, Message T behaves as a random message would, there has been observed an event, $X^2 = 155.6$, which occurs with a probability of

$$P\ (\ X^2 \geq 155.6,\ 25\ \text{d.f.}) < 0.0001.$$

Judging that such an extremely unlikely event did not occur, one rejects the random hypothesis and concludes that there has been found something which is non-random and hence causal. It is argued that this Chi-squared score, 155.6, is not of a random distribution, for it is well beyond the range of random scores, but rather is of some other type of distribution, whose range of scores is more nearly like 155.6. /This is actually the case. The Chi-squared score 155.6 is that of a transposition or plain-text distribution, whose range of Chi-squared scores is such that the average score for this size distribution is $E\ (X^2,\ \text{transposition}) = 192.4$./

The interpretation of the Chi-squared score, 32.2, for Message R is as follows. Judging that an event normally occurring about 16 percent of the time might well have occurred, one does not reject the random hypothesis, and concludes only that the Chi-squared test yields no decision in this case. In practice, one would judge that this Chi-squared score, 32.2, is of a random distribution, for there is no reason to suppose otherwise.

INTERPRETATION IN TERMS OF THE STATISTIC $\sqrt{2X^2}$ *

For message T

$$X_T^2 = 155.6$$

$$\sqrt{2X_T^2} = \sqrt{2(155.6)} = 17.64$$

$$S(\sqrt{2X_T^2}) = \sqrt{2X_T^2} - \sqrt{2(26-1)-1}$$

$$S(\sqrt{2X_T^2}) = 10.64$$

For message R

$$X_R^2 = 32.2$$

$$\sqrt{2X_R^2} = \sqrt{2(32.2)} = 8.02$$

$$S(\sqrt{2X_R^2}) = \sqrt{2X_R^2} - \sqrt{2(26-1)-1}$$

$$S(\sqrt{2X_R^2}) = 1.03$$

On reference to a table of the Normal distribution (Table 2)

$$P(\ S(\sqrt{2X^2}) \geq 10.64) < 0.0001$$

$$P(\ S(\sqrt{2X^2}) \geq 1.03\ ) \doteq 0.1515$$

The probabilities associated with the $\sqrt{2X^2}$ statistics are the same as those associated with the X^2 statistics, and for both statistics the interpretation is the same, as is given above.

* Here the statistic $\sqrt{2X^2}$ will be used even though the number of degrees of freedom $\nu = 25$ is less than 30, for the approximation is still good.

SECRET

DECLASSIFIED
NW5561+5
AUTHORITY
BY LJ NARA DATE: 4/25

NOTE. The Chi-squared test of Example 1 is equivalent to the delta IC test, in accordance with the following relation:

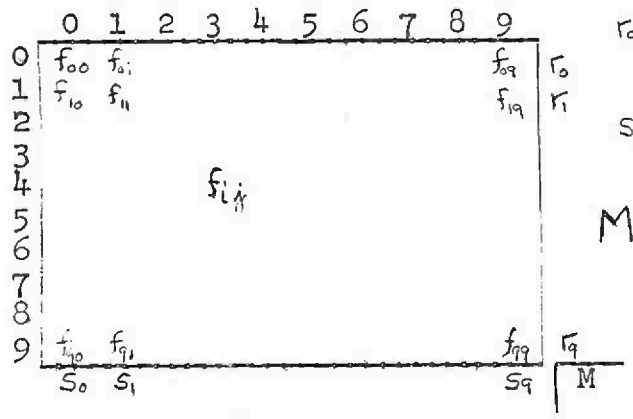
$$X^2 = (N-1)\cdot(\delta - 1) + (c-1)$$

where X^2 = Chi-squared
δ = delta IC
N = total count
c = number of categories.

EXAMPLE 2.

PROBLEM. To distinguish mononome-dinome cipher from mononomically rough, dinomically random cipher by evaluating and interpreting the Chi-squared statistic for the fit of the overall dinomic distribution to the dinomic distribution expected from a random selection from the observed mononomic distribution.

OBSERVED OVERALL DINOMIC CIPHER DISTRIBUTION.



$r_0 = \sum_{j=0}^{9} f_{0j}$ = row sum = mononome count for 0

$s_0 = \sum_{i=0}^{9} f_{i0}$ = column sum = mononome count for 0

$$M = \sum_{i=0}^{9}\sum_{j=0}^{9} f_{ij} = \sum_{i=0}^{9} r_i = \sum_{j=0}^{9} s_j$$

M = total count of dinomes

Observed frequency of dinome ij = f_{ij} $\quad i = 0,1,\ldots,9;\ j=0,1,\ldots,9.$

EXPECTED DINOMIC DISTRIBUTION FROM OBSERVED MONONOMIC DISTRIBUTION.

From the observed mononomic distribution: $r_0, r_1, \ldots, r_9$ (equivalently $s_0, s_1, \ldots, s_9$), the expected frequency of dinome ij

$$E(f_{ij}) = r_i\left(\frac{s_j}{M}\right) = s_j\left(\frac{r_i}{M}\right).$$

Thus the observed monographic roughness is expected to induce a certain amount of roughness in the digraphic distribution, even when the digraphs are selected at random. It is the purpose of this Chi-squared test to identify digraphic roughness over and above that induced by the observed monographic distribution.

DECLASSIFIED
NW5561649
AUTHORITY
BY LJ NARA DATE: 4/25

SECRET

EVALUATION OF CHI-SQUARED

$$\chi^2 = \frac{\left(f_{00} - \frac{r_0 s_0}{M}\right)^2}{\frac{r_0 s_0}{M}} + \frac{\left(f_{01} - \frac{r_0 s_1}{M}\right)^2}{\frac{r_0 s_1}{M}} + \cdots + \frac{\left(f_{0q} - \frac{r_0 s_q}{M}\right)^2}{\frac{r_0 s_q}{M}}$$

$$+ \frac{\left(f_{10} - \frac{r_1 s_0}{M}\right)^2}{\frac{r_1 s_0}{M}} + \frac{\left(f_{11} - \frac{r_1 s_1}{M}\right)^2}{\frac{r_1 s_1}{M}} + \cdots + \frac{\left(f_{1q} - \frac{r_1 s_q}{M}\right)^2}{\frac{r_1 s_q}{M}}$$

$$+ \vdots \qquad \vdots \qquad \vdots$$

$$+ \frac{\left(f_{q0} - \frac{r_q s_0}{M}\right)^2}{\frac{r_q s_0}{M}} + \frac{\left(f_{q1} - \frac{r_q s_1}{M}\right)^2}{\frac{r_q s_1}{M}} + \cdots + \frac{\left(f_{qq} - \frac{r_q s_q}{M}\right)^2}{\frac{r_q s_q}{M}} =$$

$$\chi^2 = \sum_{i=0}^{q} \sum_{j=0}^{q} \frac{\left(f_{ij} - \frac{r_i s_j}{M}\right)^2}{\frac{r_i s_j}{M}} = M\left[\left(\sum_{i=0}^{q} \sum_{j=0}^{q} \frac{f_{ij}^2}{r_i s_j}\right) - 1\right].$$

NUMERICAL EXAMPLE

Overall dinomic distributions for two messages.

Message MD

	0	1	2	3	4	5	6	7	8	9	
0	4	1	4	5	9	3	5	1	10	4	$46 = r_0$
1	6	0	3	2	4	6	0	0	9	3	$33 = r_1$
2	4	1	1	10	1	1	3	1	1	6	29
3	8	5	4	0	8	1	1	1	7	9	44
4	0	10	2	7	0	1	5	0	3	5	33
5	1	1	4	4	0	0	0	1	5	2	18
6	6	4	2	0	2	0	3	3	1	1	22
7	4	2	2	0	2	5	0	2	2	4	23
8	2	3	4	9	6	0	0	14	0	1	39
9	11	6	3	7	2	1	5	0	0	1	$36 = r_9$
	46	33	29	44	34	18	22	23	38	36	$323 = M$
	$= s_0$	$= s_1$								$= s_9$	

$$\chi^2_{MD} = 234.52$$

SECRET

DECLASSIFIED
NW556649
AUTHORITY
BY LJ NARA DATE: 4/25

~~SECRET~~

Message R

	0	1	2	3	4	5	6	7	8	9	
0	0	1	0	0	0	0	0	1	1	1	4
1	2	18	1	9	2	9	5	14	3	10	73
2	0	2	0	4	0	0	0	2	1	1	10
3	1	9	2	7	2	5	4	6	2	11	49
4	0	2	1	1	2	2	1	2	1	2	14
5	0	7	1	5	4	8	1	9	4	4	43
6	0	8	2	3	0	5	1	3	0	0	22
7	0	9	1	7	3	5	4	5	8	4	46
8	0	5	2	8	0	2	3	1	1	3	25
9	1	11	0	5	1	7	4	3	4	2	38
	4	72	10	49	14	43	23	46	25	38	324 = M

$$X^2_R = 76.28$$

INTERPRETATION. On reference to a table of the Chi-squared distribution with $81 = (10-1)^2$ degrees of freedom

MD: $P\ (X^2 \geqslant 234.52,\ 81\ df) < 0.0001$

R: $P\ (X^2 \geqslant 76.28,\ 81\ df) \doteq 0.64$

$E\ (X^2) = 81$

$\sqrt{2X^2_{MD}} = 21.65$

$S\ (\sqrt{2X^2_{MD}}) = \sqrt{2\ (234.52)} - \sqrt{2(10-1)^2 - 1}$

$S\ (\sqrt{2X^2_{MD}}) = 8.97$

$\sqrt{2X^2_R} = 12.34$

$S\ (\sqrt{2X^2_R}) = \sqrt{2\ (76.28)} - \sqrt{2\ (10-1)^2 - 1}$

$S\ (\sqrt{2X^2_R}) = -0.34$

MD: $P\ (S\ (\sqrt{2X^2}) \geqslant 8.97) < 0.0001$

R: $P\ (S\ (\sqrt{2X^2}) \geqslant -0.34) \doteq 0.63$

The Chi-squared score for Message MD is far beyond the range for random. That is, the roughness present in the dinomic distribution is well beyond that which would be induced by random selection from the observed mononomic distribution. Thus the random hypothesis is rejected. Rightly so; the message is mononome-dinome cipher.

The Chi-squared score for Message R is well within the range for random, so that the random hypothesis cannot be rejected. One would judge that this dinomic distribution has no more roughness than would be induced by random selection from the observed mononomic distribution.

NOTE. In EXAMPLE 2, CAUTION 1 was ignored without disastrous results.

~~SECRET~~

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

SECRET

EXAMPLE 3.

PROBLEM. To identify the correct alignment of two frequency distributions which are known to be slides of each other by evaluating and interpreting the Chi-squared statistic, at each alignment, for the fit of the two distributions to the distributions expected for each from a random selection from the combined distribution of the two. (It is necessary that the alphabetical sequence be known. Here it is assumed to be normal.)

TWO OBSERVED MONOGRAPHIC DISTRIBUTIONS.

i \ j	1	2	3	∘ ∘ ∘	c	
1	f_{11}	f_{12}	f_{13}	∘ ∘ ∘	f_{1c}	$r_1 = N$
2	f_{21}	f_{22}	f_{23}	∘ ∘ ∘	f_{2c}	$r_2 = N$
	s_1	s_2	s_3		s_c	$2N$

Observed frequency in distribution i of character j = f_{ij} (i=1,2; j=1,2,..., c.)

$$r_1 = \sum_{j=1}^{c} f_{1j} = N = \text{row sum} = \text{total count in distribution 1.}$$

$$r_2 = \sum_{j=1}^{c} f_{2j} = N = \text{row sum} = \text{total count in distribution 2.} *$$

$$s_1 = f_{11} + f_{21} = \text{column sum} = \text{total count of character 1.}$$

$$s_2 = f_{12} + f_{22} = \text{column sum} = \text{total count of character 2.}$$

$$\vdots \qquad\qquad \vdots$$

$$s_c = f_{1c} + f_{2c} = \text{column sum} = \text{total count of character c.}$$

$2N$ = total count in the two distributions.

EXPECTED DISTRIBUTION FOR EACH

i \ j	1	2	3	∘ ∘ ∘	c	
1	$\frac{1}{2}s_1$	$\frac{1}{2}s_2$	$\frac{1}{2}s_3$		$\frac{1}{2}s_c$	$r_1 = N$
2	$\frac{1}{2}s_1$	$\frac{1}{2}s_2$	$\frac{1}{2}s_3$		$\frac{1}{2}s_c$	$r_2 = N$
	s_1	s_2	s_3	∘ ∘ ∘	s_c	

Expected frequency in distribution i of character j

$$E(f_{ij}) = \frac{r_i s_j}{2N} = \frac{s_j}{2} = \frac{f_{1j} + f_{2j}}{2}.$$

* Distribution 2 need not have the same total count as distribution 1. With appropriate changes, the analysis is the same.

SECRET

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

~~SECRET~~

NOTE ON EXPECTED VALUES AND DEGREES OF FREEDOM.

Examples 2 and 3 involve special cases of an m x n contingency table. The general character in the observed m x n matrix is

$$f_{ij} \qquad \begin{matrix} i=1,2,\ldots, m \\ j=1,2,\ldots, n. \end{matrix}$$

The row and column sums and the total count are

$$r_i = \sum_{j=1}^{n} f_{ij} \qquad i=1,2,\ldots, m$$

$$s_j = \sum_{i=1}^{m} f_{ij} \qquad j=1,2,\ldots, n.$$

$$M = \sum_{i=1}^{m} r_i = \sum_{j=1}^{n} s_j = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}$$

The expected value of f_{ij}

$$E(f_{ij}) = \frac{r_i s_j}{M}.$$

Here $$X^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\left(f_{ij} - \frac{r_i s_j}{M}\right)^2}{\frac{r_i s_j}{M}}$$

and $E(X^2) = (m-1)\cdot(n-1) =$ number of degrees of freedom.

EVALUATION OF CHI-SQUARED.

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{c} \frac{\left(f_{ij} - \frac{f_{1j}+f_{2j}}{2}\right)^2}{\frac{f_{1j}+f_{2j}}{2}} = \sum_{j=1}^{c} \frac{(f_{1j} - f_{2j})^2}{f_{1j} + f_{2j}}$$

NUMERICAL EXAMPLE. Given two distributions aligned at two slides (one set labeled Slide C, the other Slide I) for each to evaluate and interpret the Chi-squared statistic for the fit of the two distributions to the distributions expected for each from a random selection from the combined distribution of the two.

Observed distributions: Slide C.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	TOTAL
Dist. 1	0	6	0	5	0	2	0	0	2	0	2	11	1	1	5	1	0	0	3	7	1	3	0	0	0	0	50
Dist. 2	0	5	1	2	2	4	1	1	0	0	3	6	8	2	3	2	0	0	2	3	3	2	0	0	0	0	50

~~SECRET~~

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

~~SECRET~~

Observed distributions: Slide I.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	TOTAL
Dist. 1	0	6	0	5	0	2	0	0	2	0	2	11	1	1	5	1	0	0	3	7	1	3	0	0	0	0	50
Dist. 2	5	1	2	2	4	1	1	0	0	3	6	8	2	3	2	0	0	2	3	3	2	0	0	0	0	0	50

$$X_C^2 = \frac{(0-0)^2}{0+0} + \frac{(6-5)^2}{6+5} + \cdots + \frac{(0-0)^2}{0+0}$$ (Note CAUTION 2)

$X_C^2 = 28.32$

$$X_I^2 = \frac{(0-5)^2}{0+5} + \frac{(6-1)^2}{6+1} + \cdots + \frac{(0-0)^2}{0+0}$$

$X_I^2 = 41.21.$

Interpretation. On reference to a table of the Chi-squared distribution with $25 = (26-1)\cdot(2-1)$ degrees of freedom (Table 1)

C: $0.10 < P\ (X^2 \geq 28.32,\ 25\ df) < 0.50$ from Table 1

$P\ (X^2 \geq 28.32,\ 25\ df) \doteq 0.30$ from more extensive tables

I: $0.01 < P\ (X^2 \geq 41.21,\ 25\ df) < 0.05$ from Table 1

$P\ (X^2 \geq 41.21,\ 25\ df) \doteq 0.02$ from more extensive tables

$E\ (X^2) = (26-1)\cdot(2-1) = 25$

The interpretation of the Chi-squared score 28.32 for slide C is as follows. Under the hypothesis that the two observed distributions exhibit the same highs and lows as would be expected from the combined distribution of the two (column sums), there has been observed an event normally occurring about 30 percent of the time. Such an event might well have occurred, and hence the (random) hypothesis, equivalent to assuming that the two distributions are aligned at the correct slide, can not be rejected.

The interpretation of the Chi-squared score 41.21 for slide I is as follows. Under the hypothesis that the two distributions exhibit the same highs and lows as would be expected from the combined distribution of the two, there has been observed an event which occurs only 2 percent of the time. One would judge that this is sufficiently unlikely to rule out the (random) hypothesis and conclude that the two distributions are not aligned at the correct slide.

~~SECRET~~

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

Upon consideration of all 26 slides, one would have eliminated all slides but slide *O*, the correct slide, which was not eliminated.

A tabulation of the Chi-squared scores for the correct alignment and the first 10 slides.

Slide	Chi-squared	Percent of Time	Eliminated/ Not Eliminated
0	28.32	30	NOT-E
1	41.21	2	E
2	51.58	0.12	E
3	64.10	0.003	E
4	61.38	<0.01	E
5	72.61	<0.001	E
6	65.44	<0.003	E
7	63.62	0.003	E
8	54.99	<0.06	E
9	44.53	<1.09	E
10	48.97	<0.37	E

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

APPENDIX C

TABLE I THE DELTA IC DISTRIBUTION

TABLE II THE NORMAL DISTRIBUTION

TABLE III THE CHI-SQUARED DISTRIBUTION

DECLASSIFIED
NW55649
AUTHORITY
By LJ NARA DATE: 4/25

TABLE I

PROBABILITIES RELATED TO DELTA IC SIGMAGES*

THE DELTA IC DISTRIBUTION

SIGMAGE $S(\delta)=k$	c = 10 PROBABILITY $p=P[S(\delta)\geq k]$	c = 26 PROBABILITY $p=P[S(\delta)\geq k]$	c = 30 PROBABILITY $p=P[S(\delta)\geq k]$	c = 32 PROBABILITY $p=P[S(\delta)\geq k]$
0.0	.437	.463	.465	.466
0.2	.363	.386	.388	.389
0.4	.297	.316	.318	.319
0.6	.240	.254	.255	.256
0.8	.192	.201	.202	.202
1.0	.152	.156	.156	.157
1.2	.119	.120	.119	.119
1.4	.0927	.0902	.0897	.0896
1.6	.0715	.0671	.0664	.0662
1.8	.0548	.0493	.0485	.0482
2.0	.0417	.0358	.0349	.0347
2.2	.0315	.0256	.0248	.0246
2.4	.0237	.0182	.0175	.0172
2.6	.0178	.0127	.0121	.0119
2.8	.0132	.00885	.00834	.00816
3.0	.00980	.00609	.00568	.00553
3.2	.00723	.00415	.00383	.00371
3.4	.00532	.00280	.00255	.00246
3.6	.00390	.00188	.00169	.00162
3.8	.00284	.00125	.00111	.00106
4.0	.00207	.000826	.000721	.000683
4.2	.00150	.000542	.000466	.000438
4.4	.00109	.000353	.000299	.000279
4.6	.000783	.000228	.000190	.000176
4.8	.000564	.000147	.000120	.000110
5.0	.000405	.000094	.000075	.000069
6.0	.000075	.000009	.000007	.000006
7.0	.000013	.000001	.000001	.000001
8.0	.000002			
K	p	p	p	p

* From: The Index of Coincidence, NSA Publication, January 1955,
By: H. H. Campaigne
P. 44-49

DECLASSIFIED
NW5561495
AUTHORITY
BY LJ NARA DATE: 4/25

~~SECRET~~

TABLE II

PROBABILITIES RELATED TO NORMAL SIGMAGES*

THE NORMAL DISTRIBUTION

SIGMAGE S=k	PROBABILITY P[S≥k]
0.0	.500000
0.2	.420740
0.4	.344578
0.6	.274253
0.8	.211855
1.0	.158655
1.2	.115070
1.4	.080757
1.6	.054799
1.8	.035930
2.0	.022750
2.2	.013903
2.4	.008198
2.6	.004661
2.8	.002555
3.0	.001350
3.2	.000687
3.4	.000337
3.6	.000159
3.8	.000072
4.0	.000032
4.2	.000013
4.4	.0000054
4.6	.0000021
4.8	.0000008
5.0	.0000003

* From: BIOMETRIKA TABLES FOR STATISTICIANS, Volume I
Edited by: E. S. Pearson and H. O. Hartley
P. 104-110

~~SECRET~~

SECRET

DECLASSIFIED
NW 55045
AUTHORITY
BY NARA DATE: 4

THE CHI-SQUARED DISTRIBUTION

$P(X^2 > X_0^2, \nu \text{ d.f.}) = p$

ν \ p	0.9999	0.999	0.99	0.95	0.90	0.50	0.10	0.05	0.01	0.001	0.0001
1	2×10^{-8}	2×10^{-6}	2×10^{-4}	.004	.02	.46	2.7	3.8	6.6	10.8	15.1
2	2×10^{-4}	.002	.02	.10	.21	1.4	4.6	6.0	9.2	13.8	18.4
3	.005	.02	.12	.35	.58	2.4	6.3	7.8	11.3	16.3	20.9
4	.03	.09	.30	.71	1.1	3.4	7.8	9.5	13.3	18.5	23.5
5	.08	.20	.55	1.1	1.6	4.4	9.2	11.1	15.1	20.5	25.7
6	.17	.38	.87	1.6	2.2	5.3	10.6	12.6	16.8	22.5	27.8
7	.30	.60	1.2	2.2	2.8	6.3	12.0	14.1	18.5	24.3	30
8	.47	.85	1.6	2.7	3.5	7.3	13.4	15.5	20.1	26.1	32
9	.66	1.2	2.1	3.3	4.2	8.3	14.7	16.9	21.7	27.9	34
10	.87	1.5	2.6	3.9	4.9	9.3	16.0	18.3	23.2	29.6	36
11	1.1	1.8	3.1	4.6	5.6	10.3	17.3	19.7	24.7	31	37
12	1.4	2.2	3.6	5.2	6.3	11.3	18.5	21.0	26.2	33	39
13	1.7	2.6	4.1	5.9	7.0	12.3	19.8	22.4	27.7	35	41
14	2.0	3.0	4.7	6.6	7.8	13.3	21.1	23.7	29.1	36	43
15	2.4	3.5	5.2	7.3	8.5	14.3	22.3	25.0	31	38	44
16	2.8	4.0	5.8	8.0	9.3	15.3	23.5	26.3	32	39	46
17	3.2	4.4	6.4	8.7	10.1	16.3	24.8	27.6	33	41	48
18	3.6	4.9	7.0	9.4	10.9	17.3	26.0	28.9	35	42	50
19	4.0	5.4	7.6	10.1	11.7	18.3	27.2	30	36	44	51
20	4.4	5.9	8.3	10.9	12.4	19.3	28.4	31	38	45	53
21	4.9	6.5	8.9	11.6	13.2	20.3	29.6	33	39	47	55
22	5.3	7.0	9.5	12.3	14.0	21.3	31	34	40	48	56
23	5.8	7.5	10.2	13.1	14.8	22.3	32	35	42	50	58
24	6.3	8.1	10.9	13.8	15.7	23.3	33	36	43	51	59
25	6.7	8.7	11.5	14.6	16.5	24.3	34	38	44	53	61
26	7.1	9.3	12.2	15.4	17.3	25.3	36	39	46	54	62
27	7.6	9.9	12.9	16.2	18.1	26.3	37	40	47	55	63
28	8.2	10.4	13.6	16.9	18.9	27.3	38	41	48	57	64
29	8.7	11.0	14.3	17.7	19.8	28.3	39	43	50	58	66
30	9.3	11.6	15.0	18.5	20.6	29.3	40	44	51	60	68
31	9.8	12.2	15.7	19.3	21.5	30	42	45	52	61	69
32	10.4	12.8	16.4	20.1	22.4	31	43	46	54	63	71
33	10.9	13.4	17.1	20.9	23.2	32	44	48	55	64	72
34	11.4	14.1	17.8	21.7	24.0	33	45	49	56	66	74
35	12.0	14.7	18.6	22.5	24.9	34	46	50	58	67	75

X_0^2

TABLE III

From The Chi-Squared Test, NSA Technical paper, November 1949.

SECRET

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

SECRET

APPENDIX D

FORMULAS FOR DIAGNOSTIC STATISTICS

DETAILED SIMPLIFICATION STARTING WITH DEFINITION

In this appendix are given, without explanation, the detailed algebraic steps used to change some typical diagnostic formulas into equivalent forms that are more suitable for computer applications.

1. DELTA I.C.

Distribution.

Category	1	2	3	.	.	.	c	Total
Count	f_1	f_2	f_3	.	.	.	f_c	N

$$\delta = \frac{\sum_{i=1}^{c} \frac{f_i(f_i-1)}{2}}{\frac{N(N-1)}{2} \cdot \frac{1}{c}} \qquad \text{(Definition)}$$

$$= \frac{c}{N(N-1)} \sum_{i=1}^{c} f_i(f_i-1)$$

$$= \frac{c}{N(N-1)} \left[\sum_{i=1}^{c} f_i^2 - \sum_{i=1}^{c} f_i \right]$$

$$\delta = \frac{c}{N(N-1)} \left[\left(\sum_{i=1}^{c} f_i^2 \right) - N \right] \qquad \text{(Simplified)}$$

2. CHI-SQUARED: Monographic distribution vs. uniformly flat

Distributions.

Category	1	2	3	.	.	.	c	Total
Observed	f_1	f_2	f_3	.	.	.	f_c	N

Category	1	2	3	.	.	.	c	Total
Expected	N/c	N/c	N/c	.	.	.	N/c	N

SECRET

DECLASSIFIED
NW5564S
AUTHORITY
BY LJ NARA DATE: 4/25

$$\chi^2 = \sum_{i=1}^{c} \frac{\left(f_i - N/c\right)^2}{N/c} \qquad \text{(Definition)}$$

$$= \frac{c}{N} \sum_{i=1}^{c} \left(f_i^2 - 2\frac{N}{c} f_i + \frac{N^2}{c^2} \right)$$

$$= \frac{c}{N} \left(\sum_{i=1}^{c} f_i^2 - 2\frac{N}{c} \cdot \sum_{i=1}^{c} f_i + \sum_{i=1}^{c} \frac{N^2}{c^2} \right)$$

$$= \frac{c}{N} \left(\sum_{i=1}^{c} f_i^2 - 2\frac{N}{c} \cdot N + c \cdot \frac{N^2}{c^2} \right)$$

$$= \frac{c}{N} \left(\sum_{i=1}^{c} f_i^2 - \frac{N^2}{c} \right)$$

$$\chi^2 = \frac{c}{N} \left(\sum_{i=1}^{c} f_i^2 \right) - N \qquad \text{(Simplified)}$$

χ^2 has (c-1) degrees of freedom.

THE EQUIVALENCE OF THIS CHI-SQUARED TO THE DELTA IC

$$\delta = \frac{c}{N(N-1)} \left[\left(\sum_{i=1}^{c} f_i^2 \right) - N \right]$$

$$(N-1)\delta = \frac{c}{N} \left(\sum_{i=1}^{c} f_i^2 \right) - c$$

$$(N-1)\delta = \chi^2 + N - c$$

$$(N-1)\delta = \chi^2 + N - 1 - c + 1$$

$$\chi^2 = (N-1)\cdot(\delta - 1) + (c-1).$$

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

SECRET

3. CHI-SQUARED: c x c Contingency table vs. expected from row and column sums.

Distribution.

	1	2	3	. . .	j	. . .	c	
1	f_{11}	f_{12}			f_{1j}		f_{1c}	r_1 row sums
2	f_{21}	f_{22}			f_{2j}		f_{2c}	r_2
3								
.								
.								
.								
i	f_{i1}	f_{i2}			f_{ij}		f_{ic}	r_i
.								
.								
.								
c	f_{c1}	f_{c2}			f_{cj}		f_{cc}	r_c
	S_1	S_2			S_j		S_c	N

Column sums — Total count

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{c} \frac{\left(f_{ij} - \frac{r_i s_j}{N}\right)^2}{\frac{r_i s_j}{N}} \qquad \text{(Definition)}$$

$$= \sum_{i=1}^{c} \sum_{j=1}^{c} \frac{N}{r_i s_j} \left(f_{ij}^2 - 2 \frac{r_i s_j}{N} f_{ij} + \frac{r_i^2 s_j^2}{N^2} \right)$$

$$= N \sum_{i=1}^{c} \sum_{j=1}^{c} \frac{f_{ij}^2}{r_i s_j} - 2 \sum_{i=1}^{c} \sum_{j=1}^{c} f_{ij} + \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{c} r_i s_j$$

$$= N \sum_{i=1}^{c} \sum_{j=1}^{c} \frac{f_{ij}^2}{r_i s_j} - 2N + \frac{1}{N} \cdot N^2$$

$$\chi^2 = N \left[\left(\sum_{i=1}^{c} \sum_{j=1}^{c} \frac{f_{ij}^2}{r_i s_j} \right) - 1 \right] \qquad \text{(Simplified)}$$

χ^2 has $(c-1)^2$ degrees of freedom

SECRET

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

4. CHI-SQUARED. 2 x c Contingency table vs. expected from row and column sums (where row sums are equal).

Distribution.

Category	1	2	3 . . .	j	. . .	c	Total
Observed #1	f_{11}	f_{12}		f_{1j}		f_{1c}	N row sums
Observed #2	f_{21}	f_{22}		f_{2j}		f_{2c}	N
	$f_{11}+f_{21}$ column sums			$f_{1j}+f_{2j}$		$f_{1c}+f_{2c}$	2N total count

$$\chi^2 = \sum_{j=1}^{c} \frac{\left(f_{1j} - \frac{N(f_{1j}+f_{2j})}{2N}\right)^2}{\frac{N(f_{1j}+f_{2j})}{2N}} + \sum_{j=1}^{c} \frac{\left(f_{2j} - \frac{N(f_{1j}+f_{2j})}{2N}\right)^2}{\frac{N(f_{1j}+f_{2j})}{2N}}$$

(Definition)

$$= \sum_{j=1}^{c} \frac{\left(f_{1j} - \frac{(f_{1j}+f_{2j})}{2}\right)^2}{\frac{(f_{1j}+f_{2j})}{2}} + \sum_{j=1}^{c} \frac{\left(f_{2j} - \frac{(f_{1j}+f_{2j})}{2}\right)^2}{\frac{(f_{1j}+f_{2j})}{2}}$$

$$= \sum_{j=1}^{c} \frac{\left(\frac{f_{1j}-f_{2j}}{2}\right)^2}{\left(\frac{f_{1j}+f_{2j}}{2}\right)} + \sum_{j=1}^{c} \frac{\left(\frac{f_{2j}-f_{1j}}{2}\right)^2}{\left(\frac{f_{1j}+f_{2j}}{2}\right)}$$

$$= 2\sum_{j=1}^{c} \frac{\left(\frac{f_{1j}-f_{2j}}{2}\right)^2}{\left(\frac{f_{1j}+f_{2j}}{2}\right)}$$

$$\chi^2 = \sum_{j=1}^{c} \frac{(f_{1j}-f_{2j})^2}{(f_{1j}+f_{2j})}$$

(Simplified)

χ^2 has $(2-1)\cdot(c-1) = (c-1)$ degrees of freedom.

DECLASSIFIED
NW55649
AUTHORITY
BY NARA DATE: 4/25

BIBLIOGRAPHY.

1. Oyer, Paul D. and Hurley, G., "Supersteth Routines and Statistics (Revised)", PROD-03 Informal No. 12, 10 June 1957 (SECRET)

2. Campaigne, H. H., "The Index of Coincidence", NSA Publication, 12 January 1955. (CONFIDENTIAL)

3. Greenwood, R. E., and Lotz, W., "The Chi-Square Test", AFSA-34, 14 March 1952. (CONFIDENTIAL)

4. Kupperman, M. "On Comparing Two Observed Frequency Counts", The NSA Technical Journal, Vol. III, No. 2 (1958), pp. 69-75. (SECRET)

5. Leahy, F. T. "Notes on Chi-Square in Cryptanalysis", The NSA Journal, Vol. II, No. 3 (1957), pp. 61-72 (SECRET)

6. Oyer, Paul D., "The STETHOSCOPE Program", PROD-03 Informal No. 19 (forthcoming).

Dear Patron:

We regret that the enclosed photocopies are the best we were able to obtain using our normal reproduction process. This is caused primarily by the age and faded conditions of some of the documents from which these copies are made.

BEST AVAILABLE COPY

DECLASSIFIED
NW5516449
AUTHORITY
BY NARA DATE: 4/25

TOP SECRET EIDER

RESULTS OF DIAGNOSTIC TESTS ON:

1. Total number of messages _____

2. Batching

   Batch (a) Definition _____ Messages

   Batch (b) Definition _____ Messages

   Batch (c) Definition _____ Messages

   Batch (d) Definition _____ Messages

3. Observations of Message length counts from:

   (i) Overall _____

   (ii) Batch (a) _____

   (iii) Batch (b) _____

   (iv) Batch (c) _____

   (v) Batch (d) _____

4. Observations from examination of write outs of A1-A10 and Z3-Z0:
   (Tests performed for each position: Monome counts, doublet counts, delta counts, digit r.r. counts).

   (i) Overall

   (ii) Batch (a)

   (iii) Batch (b)

   (iv) Batch (c)

   (v) Batch (d)

   (vi) Any other observations

5. Results of various counts of cipher text from at least 3 messages in each batch

   (i) Batch (a) Monome

   Dinome on both cuts

   Trinome cut including A1
   cut excluding A1

TOP SECRET EIDER

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

Tetranome cut including Al
cut excluding Al

Pentanome (Group)

Delta Count

Digit "e" of group

Digit "a" of group

(ii) Batch (b) Monome

Dinome on both cuts

Trinome cut including Al
cut excluding Al

Tetranome cut including Al
cut excluding Al

Pentanome (Group)

Delta Count

Digit "e" of group

Digit "a" of group

(iii) Batch (c) Monome

Dinome on both cuts

Trinome cut including Al
cut excluding Al

Tetranome cut including Al
cut excluding Al

Pentanome (Group)

Delta Count

Digit "e" of group

Digit "a" of group

(iv) Batch (d) Monome

Dinome on both cuts

DECLASSIFIED.
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

Trinome cut including A1
cut excluding A1

Tetranome cut including A1
cut excluding A1

Pentanome (Group)

Delta Count

Digit "e" of group

Digit "a" of group

6. Observations from subtracting A1 "a b c d e" from A2a

7. Results of group for group examination of re-sends

8. Results of examination of filler digits on assumption of 4f code with I.N.T.

9. Any other unusual features noted:

10. General conclusions from examination

DECLASSIFIED
NW55645
AUTHORITY
BY LJ NARA DATE: 4/25

~~TOP SECRET EIDER~~

REPORTING FORM, UNSOLVED SYSTEMS

1. Title _____ 2. Service description _____

3. Time element covered _____ to _____

4. Approximate volume _____ per month _____ on hand

5. Elements in _____ letter groups? _____ digit groups?

(If letters, how many in cipher alphabet _____ Missing elements, if any?

_____

6. Non-textual groups, if any _____ Indicator if any _____

7. Probable indicator explanation _____

8. Evidence of tailing or trailing? _____

9. Phenomena noted? _____

10. Length test. No. of messages _____ : Remainder test, div. by 2 _____

Div. by 3 _____ Div. by 4 _____ Comments _____

11. Statistical tests:

a. Monographic: Message No. _____ _____ Elements counted _____ I.C.
No. _____ _____ " " _____ I.C.
No. _____ _____ " " _____ I.C.
No. _____ _____ " " _____ I.C.
No. _____ _____ " " _____ I.C.

Combined count: _____ Elements counted _____ I.C.

b. Digraphic I.C. Message No. _____ _____ Elements counted _____ I.C.
No. _____ _____ " " _____ I.C.
No. _____ _____ _____ I.C.

Combined count: _____ Elements counted _____ I.C.

c. Results of width tests, if any

Width _____, Average Depth _____, I.C. _____

Width _____, Average Depth _____, I.C. _____

Width _____, Average Depth _____, I.C. _____

~~TOP SECRET EIDER~~

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

d. Evidence of local roughness? Coincidence at offset of

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26

e. Positional roughness? IC of positions 1____ 2____ 3____ 4____ 5____

f. Running delta, normal components? Results if any _____

12. Repeats noted? At beginnings? _____ At ends? _____ In text? _____

13. Overlaps? _____ Depths available up to _____ deep. Length of unit_____

Comparisons counted? _____ I.C. _____

14. Bust messages? _____

15. Any related facts, or collateral information, such as previous systems used, probable type of system, etc.

_____

_____

16. On operating level, who is most familiar with system? _____

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

~~SECRET~~

SUPERSTETH PRESCRIPTIONS

27 October, 1958

SUPPLEMENT TO PROD-03 INFORMAL NO. 18

SUPERSTETH DIAGNOSTIC PROGRAMS, dated 25 August 1958*

*Errata - the date was erroneously printed as 1957

DISTRIBUTION:

D/D	1	ADVA-43	2
SPEC. ASS'T.	1	ADVA-44	2
BLO	13	ACOM	1
USLO	5	ACOM-03	1
PROD	1	ACOM-1	2
PROD-03	315 (reserve)	ACOM-2	2
MPRO	1	ACOM-22	2
MPRO-03	6	ACOM-23	3
MPRO-1	13	ACOM-24	1
MPRO-3	11	ACOM-4	3
MPRO-5	18	ALLO	1
CREF	10	ALLO-03	1
CREF-407	1	ALLO-1	5
COLL-03	1	ALLO-2	5
GENS	1	ALLO-303	2
GENS-04	1	ALLO-31	1
GENS-14	2	ALLO-32	1
GENS-22	2	ALLO-33	2
GENS-35	2	ALLO-4	1
GENS-403	1	ALLO-403	2
GENS-43	4	ALLO-41	3
GENS-503	5	ALLO-42	1
ADVA	1	R/D	1
ADVA-03	1	ANEQ	6
ADVA-1	1	ENGR	1
ADVA-12	1	REMP	1
ADVA-13	2	REMP-1	2
ADVA-14	2	REMP-102	2
ADVA-2	1	DOC-2	1
ADVA-21	2	DOC-202	2
ADVA-22	3	STED-4	2
ADVA-23	1	TNG	1
ADVA-31	3	TNG-03	2
ADVA-4	1	TNG-2	1
		TNG-3	1

This supplement is to be attached to PROD-03 Informal No. 18.

~~SECRET~~

DECLASSIFIED
NW55645
AUTHORITY
BY LJ NARA DATE: 4/25

SECRET

SUPERSTETH PRESCRIPTIONS

L-1 Literal Traffic. Nothing known or assumed

TESTS

1.1	Mono count
1.2.1	Mono IC and sigmage
1.2.2	Mono chi-square and sigmage
1.4.1	Mono distribution by position in group--group sizes 2, 3, 4, and 5
1.4.2	Mono IC and sigmage on mono distribution by position in group--group sizes 2, 3, 4, and 5
1.4.2	Chi-square and sigmage on mono distribution by position in group--group sizes 2, 3, 4, and 5
1.4.3	Total chi-square and sigmage for each group size
1.4.4	Chi-square and sigmage between mono distributions by position in group--all combinations of positions, slide of 0, group sizes 2, 3, 4, and 5
2.1.1*	Digraphic count
2.1.2*	Digraphic IC and sigmage
2.1.3*	Digraphic chi-square and sigmage
2.2.1*	Chi-square and sigmage on the row totals
2.2.2*	Chi-square and sigmage on the column totals
2.2.3*	Chi-square and sigmage on the row totals versus the column totals
2.3.1*	Chi-square and sigmage for rutting
2.3.2*	Chi-square and sigmage for symmetry
2.3.3*	Chi-square and sigmage for bias
2.3.4*	Chi-square and sigmage on all other diagonals
2.4.1*	Poisson printout for digraphs
2.4.2*	High frequency digraph printout
3.1	Mono local roughness on slides 1 through 33
4.1	Polygraphic IC and sigmage up to a calculated size
4.2	Index of cohesion for each sized n-graph of 4.1
4.3	Printout of repeated text of a calculated length along with positions and offsets
4.4	Sum check between repeats
5.1	Average of column IC's and root mean square sigmage on sponsor-designated widths
5.2	Mono IC and sigmage on individual columns that are over 3 sigmas rough for widths designated in 5.1

* These tests are run on overall, on-cut, and off-cut digraphs

SECRET

DECLASSIFIED
NW5566495
AUTHORITY
BY LJ NARA DATE: 4/25

D-1 Digital Traffic. Nothing known or assumed

TESTS

1.1	Mono count
1.1.1	Mono count in frequency order
1.2.1	Mono IC and sigmage
1.2.2	Mono chi-square and sigmage
1.4.1	Mono distribution by position in group--group sizes 2, 3, 4, and 5
1.4.2	Mono IC and sigmage on mono distribution by position in group--group sizes 2, 3, 4, and 5
1.4.2	Chi-square and sigmage on mono distribution by position in group--group sizes 2, 3, 4, and 5
1.4.3	Total chi-square and sigmage for each group size
1.4.4	Chi-square and sigmage between mono distributions by position in group--all combinations of positions, slide of 0, group sizes 2, 3, 4, and 5.
2.1.1	Digraphic count--overall, on-cut, and off-cut
2.1.2	Digraphic IC and sigmage--overall, on-cut, and off-cut
2.1.3	Digraphic chi-square and sigmage--overall, on-cut, and off-cut
2.2.1	Chi-square and sigmage on the row totals--overall, on-cut, and off-cut
2.2.2	Chi-square and sigmage on the column totals--overall, on-cut, and off-cut
2.2.3	Chi-square and sigmage on row totals versus column totals--overall, on cut, and off-cut
2.3.1	Chi-square and sigmage for rutting--overall, on-cut, and off-cut
2.3.2	Chi-square and sigmage for symmetry-- " " and "
2.3.3	Chi-square and sigmage for bias-- " " and "
2.3.4	Chi-square and sigmage on all other diagonals--overall, on-cut, and off-cut
2.4.1	Poisson printout for digraphs--overall, on-cut, and off-cut
2.4.2	High frequency digraph printout-- " " and "
3.1	Mono local roughness on slides 2 through 33
3.2	Mono local coincidences by position in group--group sizes 2, 3, 4, 5, and 6
4.1	Polygraphic IC and sigmage up to a calculated size
4.2	Index of cohesion for each sized n-graph of 4.1
4.3	Printout of repeated text of a calculated length along with positions, and offsets
4.4	Sum check between repeats

DECLASSIFIED
NW5561/49
AUTHORITY
BY LJ NARA DATE: 4/25

DL-2 Digital or literal traffic suspected to be dinomic or digraphic

TESTS

1.1	Mono count
1.1.1	Mono count in frequency order
1.2.1	Mono IC and sigmage
1.2.2	Mono chi-square and sigmage
1.4.1	Mono distribution by position in group--group size 2
1.4.2	Mono IC and sigmage on mono distributions by position in group--group size 2
1.4.2	Chi-square and sigmage on mono distributions by position in group--group size 2
1.4.3	Total chi-square and sigmage for group size 2
1.4.4	Chi-square and sigmage between mono distributions by position in group--slides 1 through 2, group size 2
2.1.1*	Digraphic count
2.1.2*	Digraphic IC and sigmage
2.1.3*	Chi-square on the digraphic contingency table
2.2.1*	Chi-square and sigmage on the row totals
2.2.2*	Chi-square and sigmage on the column totals
2.2.3*	Chi-square and sigmage on the row totals versus the column totals
2.3.1*	Chi-square and sigmage for rutting
2.3.2*	Chi-square and sigmage for symmetry
2.3.3*	Chi-square and sigmage for bias
2.3.4*	Chi-square and sigmage for all other diagonals
2.4.1*	Poisson printout for digraphs
2.4.2*	High frequency digraph printout
4.1	Polygraphic IC and sigmage up to a calculated size
4.2	Index of cohesion for each sized n-graph of 4.1
4.3	Printout of repeated text of a calculated length along with positions and offsets
4.4	Sum check between repeats

* These tests are run on overall, on-cut, and off-cut digraphs

DECLASSIFIED
NW55649
AUTHORITY
BY NARA DATE: 4/25

DL-3 Digital or literal traffic suspected to be trinomic or trigraphic

TESTS

1.1	Mono count
1.2.1	Mono IC and sigmage
1.4.1	Mono distribution by position in group--group size 3
1.4.2	Mono IC and sigmage on mono distribution by position in group--group size 3
1.4.2	Chi-square and sigmage on mono distribution by position in group--group size 3
1.4.3	Total chi-square and sigmage for group size 3
1.4.4	Chi-square and sigmage between mono distributions by position in group--all combinations of positions, slides of 1 through C, group size 3
3.1	Mono local roughness on slides 1 through 33
3.2	Mono local coincidences by position in group--group sizes 2, 3, 4, 5, and 6.
4.1	Polygraphic IC and sigmage up to a calculated size
4.2	Index of cohesion for each sized n-graph of 4.1
4.3	Printout of repeated text of a calculated length along with positions and offsets
4.4	Sum check between repeats
2.5	Digraph within group--group size 3
2.1.1	Digraphic count
2.1.2	Digraphic IC and sigmage
2.1.3	Digraphic chi-square and sigmage
2.4.1	Poisson printout for digraphs
5.1	Average of column IC's and root mean square sigmage for widths 3, 6, 9, ... , 78,81.
5.2	Mono IC and sigmage on individual columns that are over 3 sigmas rough for widths 3, 6, 9, ... , 78, 81.

DECLASSIFIED
NW55649
AUTHORITY
BY NARA DATE: 4/25

SECRET

DL-4 Digital or literal traffic suspected to be tetranomic

TESTS

1.1 Mono count
1.2.1 Mono IC and sigmage
1.2.2 Mono chi-square and sigmage
1.4.1 Mono distribution by position in group--group size 4
1.4.2 Mono IC and sigmage on mono distribution by position in group--group size 4
1.4.2 Chi-square and sigmage on mono distribution by position in group--group size 4
1.4.3 Total chi-square and sigmage for group size 4

2.5 Digraphs within group--group size 4
- 2.1.2 Digraphic IC and sigmage
- 2.1.3 Digraphic chi-square and sigmage

4.1 Polygraphic IC and sigmage up to a calculated size
4.2 Index of cohesion for each sized n-graph of 4.1
4.3 Printout of repeated text of a calculated length along with positions and offsets
4.4 Sum check between repeats

5.1 Average of column IC's and root mean square sigmage for widths 4, 8, 12, ... , 76, 80
5.2 Mono IC and sigmage on individual columns that are over 3 sigmas rough for widths 4, 8, 12, ... , 76, 80

SECRET

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

DL-5 Digital or literal traffic suspected to be pentanomic or pentagraphic

TESTS

1.1	Mono count
1.1.1	Mono count in frequency order
1.2.1	Mono IC and sigmage
1.2.2	Mono chi-square and sigmage
1.4.1	Mono distributions by position in group--group size 5
1.4.2	Mono IC and sigmage on mono distributions by position in group--group size 5
1.4.2	Chi-square and sigmage on mono distributions by position in group--group size 5
1.4.3	Total chi-square and sigmage for group size 5
1.4.4	Chi-square and sigmage between mono distributions by position in group--all combinations of positions, slides 0 through 10, group size 5
2.5	Digraphs within group--group size 5
	2.1.1 Digraph count
	2.1.2 Digraphic IC and sigmage
	2.1.3 Chi-square on the digraphic contingency table
	2.4.1 Poisson printout for digraphs
3.1	Mono local roughness on slides 1 through 33
3.2	Mono local coincidences by position in group--group sizes 2, 3, 4, 5, and 6
4.1	Polygraphic IC and sigmage up to a calculated size
4.2	Index of cohesion for each sized n-graph
4.3	Printout of repeated text of a calculated length along with positions and offsets
4.4	Sum check between repeats
5.1	Average of column IC's and root mean square sigmage on widths 5, 10, ... , 75, 80
5.2	Mono IC and sigmage on individual columns that are over 3 sigmas rough for widths 5, 10, 15, ... , 75, 80

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

DL-Digraphic. Suspected digraphic. Literal or digital traffic.

TESTS

(all tests listed below are run on overall, on-cut and off-cuts of message)

2.1.1	Digraphic count
2.1.2	Digraphic IC and sigmage
2.1.3	Digraphic chi-square and sigmage
2.2.1	Chi-square and sigmage on the row totals
2.2.2	Chi-square and sigmage on the column totals
2.2.3	Chi-square and sigmage on row totals versus column totals on slides 1 through C.
2.3.1	Chi-square and sigmage on rutting
2.3.2	Chi-square and sigmage on symmetry
2.3.3	Chi-square and sigmage on bias
2.3.4	Chi-square and sigmage on all diagonals
2.4.1	Poisson printout for digraphs
2.4.2	High frequency digraph printout

DL-Special Hagelin. Hagelin machine or suspected cycles. Available for digital as well as literal

TESTS

1.1	Mono count
1.2.1	Mono IC and sigmage
5.1	Average of column IC's and root mean square sigmage for widths 2 through 50
5.2	Mono IC and sigmage on individual columns that are over 3 sigmas rough for widths 2 through 50

DL-Super widths. Suspected cycles on digital or literal

TESTS

1.1	Mono count
1.2.1	Mono IC and sigmage
5.1	Average of column IC's and root mean square sigmage for widths 2 through 128
5.2	Mono IC and sigmage on individual columns that are over 3 sigmas rough for widths 2 through 128

DECLASSIFIED
NW5564S
AUTHORITY
BY LJ NARA DATE: 4/25

DL-4A Digital or literal traffic suspected of being tetragraphic. Two tests are changed from DL-4

TESTS

1.1	Mono count
1.2.1	Mono IC and sigmage
1.4.1	Mono distribution by position in group--group size 4
1.4.2	Mono IC and sigmage on mono distribution by position in group--group size 4
1.4.2	Chi-square and sigmage on mono distribution by position in group--group size 4
1.4.3	Total chi-square and sigmage for group size 4
1.4.4	Chi-square and sigmage between mono distributions by position in group--all combinations of positions, slides 1 through c, group size 4
3.1	Mono local roughness on slides 2 through 33
3.2	Mono local coincidences by position in group--group sizes 2, 3, 4, 5, and 6.
4.1	Polygraphic IC and sigmage up to a calculated size
4.2	Index of cohesion for each sized n-graph of 4.1
4.3	Printout of repeated text of a calculated length along with positions and offsets
4.4	Sum check between repeats
5.1	Average of column IC's and root mean square sigmage for widths 4, 8, 12, ... , 76, 80.
5.2	Mono IC and sigmage on individual columns that are over 3 sigmas rough for widths 4, 8, ... , 76, 80

DL-Cycle Search. A minimum type test looking for cycles

TESTS

1.1	Mono count
4.1	Polygraphic IC and sigmage up to a calculated size
4.2	Index of cohision for each sized n-graph of 4.1
4.3	Printout of repeated text of a calculated length along with positions and offsets
4.4	Sum check
5.1	Average of column IC's and root mean square sigmage for widths specified by the sponsor
5.2	Mono IC and sigmage on individual columns that are over 3 sigmas rough for widths in 5.1

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

~~SECRET~~

PROD-03

0492

INFORMAL NO. 12
10 June 1957

INFORMAL

SUBJECT: SUPERSTETH ROUTINES AND STATISTICS (REVISED)

PREPARED BY: Paul D. Oyer
George Hurley

DISTRIBUTION:

PROD-03	2	
PROD-03M	2	
PROD-03	45	(reserve)
MATH	2	
DOC-2	2	
ANEQ	6	
BLO	13	
USLO	2	
MPRO-03	3	
MPRO-05	6	
MPRO-1	2	
MPRO-23	2	
MPRO-3	2	
MPRO-5	24	
STED-4	1	
ADVA-31	1	
TOTAL	115	

~~SECRET~~

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

~~SECRET~~

SUPERSTETH

Reprint 31 December 1957

This informal, published in the form of a request for the writing of the SUPERSTETH "general purpose program", is being reprinted to satisfy increasing sponsors' demands.

The SUPERSTETH collection of diagnostic tests is now available for diagnosis of cipher or key streams. Because of its versatility, it might even have other applications.

Hopefully the results (statistics and significance tests) of SUPERSTETH runs, when properly interpreted, may lead to the discovery of the method which caused the observed phenomena.

Requests for SUPERSTETH runs will be made by checking selections on the REBUS. To aid the sponsor, several sets of standard selections (called Prescriptions), have already been made for certain general types of systems.

Sponsors are invited to obtain assistance from PROD-03 or MPRO-05 in requesting SUPERSTETH runs and in interpretation of results from both a statistical and cryptanalytic point of view.

All routines are operational except those listed below. The stages of development and time estimates are as follows:

1. Being debugged or written. (Expected 11 January 1958 or sooner)

   3.1 3.2 3.3

2. Not yet written. (Expected 11 March 1958 or sooner)

   1.3 5.3 5.4

-0-

~~SECRET~~

DECLASSIFIED
NW5561149
AUTHORITY
BY LJ NARA DATE: 4/25

SECRET

SUPERSTETH ROUTINES AND STATISTICS (REVISED)

This paper, which supersedes PROD-03 Informal No. 7 of 14 January 1957, contains some additions and corrections to that paper. It reflects helpful constructive criticisms and suggestions by other components of NSA and by GCHQ.

The SUPERSTETH routines, aimed primarily toward stream analysis, are recommended (either individually or packaged) as useful additions to any computer library.

It is hoped that the final SUPERSTETH program will have even greater capacities and flexibilities than those requested in this paper, but it is recognized that some curtailment may be necessary.

On the complex routines, programmers may find it feasible to write separate routines for inputs of size 2000, 5000, and 12,000 for optimum running-time efficiency. Experience with STETHOSCOPE and LULU suggests that 90 per cent of job requests (for single stream analysis) are for size 2000 or less.

MPRO-503 has prepared a REBUS by means of which the sponsor may easily and clearly select the SUPERSTETH routines desired. This appears in Appendix A.

Samples of recommended output formats, appearing in Appendix B, also have been prepared by MPRO-503. In order to assure more efficient use of output, it is hoped that programmers will provide for output as nearly like these formats as is feasible.

General Requirements:

Provide for a maximum stream length of about 12,000 characters and for any alphabet size from 2 to 64.

Make allowance for alteration of thresholds and parameters in accordance with sponsor's requests.

Make allowance for dits.

Indicate negative sigmages with a minus sign.

Occasionally in computing certain Chi-squared statistics, a term in the sum may have its denominator equal to zero. In all such cases the numerator of that term will also be zero. Assign the value one (1) to that term and compute the entire statistic as required.

Allow sponsor to specify the order of his alphabet. For example, a 26-letter alphabet could be used in the order A X F W G ... Z. The specified order should then be retained for all the tests on this stream. (As in print-out of monographic counts, alphabet sliding, row and column coordinates of a digraphic matrix, delta stream, etc.)

Sponsor should also be allowed to choose the baud configurations or codings (up to 6 levels) to be used for his stream.

DECLASSIFIED
NW5564S
AUTHORITY
BY NARA DATE: 4/25

The output (and input) symbols to be used with alphabets of size 2 to 36 (which are not in Cyrillic characters) should be selected from the 26 letters A to Z plus the digits Ø to 9. Additional symbols for alphabets of sizes 37 to 64 should contain all the above symbols plus the symbols 37-, 38-, ..., 64-. (The numbers 37 to 64 followed by a dash.)

For Cyrillic alphabets, the Cyrillic characters may be used throughout the printout. In this case, the routine labels on the output will read СУПЕРСТЕТХ for SUPERSTETH, etc.

Make provision to print <u>two</u> copies of all SUPERSTETH runs.

Arrange for printout on $8\frac{1}{2}$ by 11 fanfold paper. Start the printout for each sub-routine on a new page. On each page print SUPERSTETH, the number of the run, sponsor's stream identification(s), and the sub-routine identification. (See Appendix B)

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

SUB-ROUTINES

1. MONOGRAPHIC TESTS

1.1 Print the monographic distribution in the alphabetical order as specified by sponsor. (In the REBUS, the sponsor may label the alphabet with Roman or Cyrillic characters, digits, or Baudot configurations using an arbitrary alphabetical ordering). Identify the count for each character with the sponsor's label. Also print the total number of valid characters and the number of dits.

1.1.1 Print the distribution in frequency order, with the most frequent character first.

1.2 STATISTICAL TESTS

1.2.1 Print the monographic delta I.C., δ , and its sigmage, S (δ).

$$\delta = \frac{c}{N(N-1)}\left[\left(\sum_{i=1}^{c} fi^2\right) - N\right]$$

$$S(\delta) = \frac{N(\delta-1)}{\sqrt{2(c-1)}}$$

1.2.2 Print the Chi-Squared, χ^2, and its (approximate) sigmage, S ($\sqrt{2\chi^2}$).

$$\chi^2 = \left(\frac{c}{N}\sum_{i=1}^{c} fi^2\right) - N$$

$$S(\sqrt{2\chi^2}) = \sqrt{2\chi^2} - \sqrt{2(c-1)-1}$$

Where: c = number of categories or alphabet size
fi = total number of valid characters in category i.
$N = \sum_{i=1}^{c} fi$ = total number of valid characters
(a dit is not a valid character.)

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

1.3 **LEVEL COUNTS.** (For baudot-type codings)

Print the distribution of the dots and crosses, and the sigmage, S_ℓ, for each level.

Is p_ℓ specified by the sponsor??

$$S_\ell = \frac{(n_\ell(x) - N p_\ell)}{\sqrt{N p_\ell (1-p_\ell)}}$$

Where: $n_\ell(x)$ = number of crosses on level ℓ.

p_ℓ = probability of a cross on level ℓ. (This need not be $\frac{1}{2}$ as, for example, in a 26 letter, 5-level alphabet.)

1.4 **MONOGRAPHS BY POSITION-IN-GROUP.**

1.4.1 For selected group sizes 2 to 6, considering the stream as written on a width equal to the group size, count and print out the monographic distribution of the characters in each position of the group (column).

1.4.2 For the distribution from each position of the group (for selected group sizes 2 to 6), print out the delta I.C., the sigmage of delta I.C., Chi-square, and the sigmage of Chi-square.

$$\delta_r = \frac{c}{N_r(N_r-1)}\left[\left(\sum_{i=1}^{c} f_{ir}^2\right) - N_r\right]$$

$$r = 1, 2, \ldots, g$$

$$S(\delta_r) = \frac{N_r(\delta_r - \delta)}{\sqrt{2(c-1)}}$$

$$\chi_r^2 = \left[\frac{N}{N_r}\left(\sum_{i=1}^{c} \frac{f_{ir}^2}{f_i}\right)\right] - N_r$$

$$S(\sqrt{2\chi_r^2}) = \sqrt{2\chi_r^2} - \sqrt{\frac{2(g-1)(c-1)}{g} - 1}$$

DECLASSIFIED
NW5556495
AUTHORITY
BY LL NARA DATE: 4/25

Where: f_{ir} = the number of valid characters in position r which fall in category i.

$N_r = \sum_{i=1}^{c} f_{ir}$ = total number of valid characters in position r.

$N = \sum_{i=1}^{c} f_i = \sum_{r=1}^{q} N_r$ = total number of valid characters in the stream.

δ = overall monographic delta I.C. of 1.2.1.

1.4.3 For each selected group size print out the total Chi-squared, $\hat{\chi}^2$, and the sigmage, S $(\sqrt{2\hat{\chi}^2})$.

$$\hat{\chi}^2 = \sum_{r=1}^{q} \chi_r^2$$

$$S(\sqrt{2\hat{\chi}^2}) = \sqrt{2\hat{\chi}^2} - \sqrt{2(q-1)(c-1)-1}$$

1.4.4 For the selected group sizes 2 to 6, compare the distributions of the respective pairs of positions, taken in all combinations, by means of the statistic χ_{rs}^2. Print out the value of this statistic and the sigmage S $(\sqrt{2\chi_{rs}^2})$.

$$\chi_{rs}^2 = \frac{N_{rs}^2}{N_r \cdot N_s}\left[\left(\sum_{i=1}^{c} \frac{f_{ir}^2}{f_{ir}+f_{is}}\right) - \frac{N_r^2}{N_{rs}}\right]$$

$$S(\sqrt{2\chi_{rs}^2}) = \sqrt{2\chi_{rs}^2} - \sqrt{2(c-1)-1}$$

$$(r, s = 1, 2, \cdots, q\,;\ r < s)$$

Where: $N_{rs} = N_r + N_s$ = the total number of valid characters in positions r and s.

f_{is} = the number of valid characters in position s which fall in category i.

DECLASSIFIED
NW556459
AUTHORITY
BY LJ NARA DATE: 4/25

1.4.5 For the selected group sizes 2 to 6, compare the distributions of the respective pairs of positions (in all combinations) at selected slides, i, from 1 to (C-1) by the statistic $_i\chi^2_{rs}$. Print out the value of this statistic and the sigmage.

$$_i\chi^2_{rs} = \sum_{i=1}^{C} \left[\frac{(f_{ir} - f_{(i+j) \bmod c, s})^2}{f_{ir} + f_{(i+j) \bmod c, s}} \right]$$

$$(r, s = 1, 2, \cdots, g \; ; \; r < s)$$

$$S\left(\sqrt{2\,{}_i\chi^2_{rs}}\right) = \sqrt{2\,{}_i\chi^2_{rs}} - \sqrt{2(C-1)-1}$$

DECLASSIFIED
NW55645
AUTHORITY
BY LJ NARA DATE: 4/25

2. DIGRAPHIC TESTS

2.1 Digraphic distributions and statistics.

2.1.1 Print out the (selected) .1 Overall
.2 On-cut
.3 Off-cut
digraphic distributions of the stream, showing row and column totals.

2.1.2 Print out the delta I.C. and the sigmage of delta I.C. for the (selected) .1 Overall
.2 On-cut
.3 Off-cut
digraphic distributions. If the sigmage exceeds 3.ø (or some other specified value), print the distribution (if it has not already been printed).

$$\delta_2 = \frac{c^2}{M(M-1)}\left[\left(\sum_{i=1}^{c}\sum_{j=1}^{c} f_{ij}^2\right) - M\right]$$

$$S(\delta_2) = \frac{M(\delta_2 - 1)}{\sqrt{2c(c+2)}}\text{, for overall.}$$

$$S(\delta_2) = \frac{M(\delta_2 - 1)}{\sqrt{2(c^2-1)}}\text{, for on-cut and off-cut.}$$

Where: M = total number of valid digraphs in this distribution (either overall, on-cut, or off-cut).
(A valid digraph consists of two valid characters).

f_{ij} = the number of valid digraphs whose first character is in category i and whose second character is in category j .

DECLASSIFIED
NW55645
AUTHORITY
BY LJ NARA DATE: 4/25

2.1.3 Print the Chi-squared and sigmage, S ($\sqrt{2\chi^2}$), for the (selected) .1 Overall
.2 On-cut
.3 Off-cut
digraphic distributions. If any sigmage exceeds 3.ϕ (or other specified value) print the associated digraphic distribution (unless it has already been printed).

$$\chi^2 = M\left[\left(\sum_{i=1}^{c}\sum_{j=1}^{c}\frac{f_{ij}^2}{r_i s_j}\right) - 1\right]$$

$$S(\sqrt{2\chi^2}) = \sqrt{2\chi^2} - \sqrt{2(c-1)^2 - 1}$$

Where: $r_i = \sum_{j=1}^{c} f_{ij}$ = total number of valid characters in row i.

$s_j = \sum_{i=1}^{c} f_{ij}$ = total number of valid characters in column j.

2.2 Row and Column Statistics

For the (selected) .1 Overall
.2 On-cut
.3 Off-cut
digraphic distributions, print the following Chi-squared statistics and their sigmages. If any sigmage exceeds 3.ϕ (or other specified value), print the associated distribution (unless printed above).

2.2.1 Chi-Square of Row Totals

$$\chi_r^2 = \left(\frac{c}{M}\sum_{i=1}^{c} r_i^2\right) - M$$

$$S(\sqrt{2\chi_r^2}) = \sqrt{2\chi_r^2} - \sqrt{2(c-1)-1}$$

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

2.2.2 Chi-Square of Column Totals

$$\chi^2_s = \left(\frac{c}{M}\sum_{j=1}^{c} s_j^2\right) - M$$

$$S\left(\sqrt{2\chi^2_s}\right) = \sqrt{2\chi^2_s} - \sqrt{2(c-1)-1}$$

2.2.3 Rows versus Columns (slide zero)

$$\chi^2 = \sum_{i=1}^{c}\left[\frac{(r_i - s_i)^2}{r_i + s_i}\right]$$

$$S\left(\sqrt{2\chi^2}\right) = \sqrt{2\chi^2} - \sqrt{2(c-1)-1}$$

where M, r_i and s_j are defined in 2.1.

2.2.4 For the (selected) .1 Overall
.2 On-cut
.3 Off-cut
digraphic counts, compare the distributions of row and column totals at selected slides, j, from 1 to $(c-1)$ by means of the Chi-Squared and sigmage.

$$_j\chi^2 = \sum_{i=1}^{c}\left[\frac{(r_i - s_{(i+j) \bmod c})^2}{r_i + s_{(i+j) \bmod c}}\right]$$

$$S\left(\sqrt{2_j\chi^2}\right) = \sqrt{2_j\chi^2} - \sqrt{2(c-1)-1}$$

for $j = 1, 2, \ldots, (c-1)$.

2.3 DIAGONAL STATISTICS.

For the (selected) .1 Overall
.2 On-cut
.3 Off-cut
digraphic distributions, print the following Chi-squared statistics and their sigmages. If any sigmage exceeds 3.ϕ (or other specified value), print the associated distribution (unless printed above).

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

2.3.1 The Chi-squared for Rutting (main diagonal).

$$\chi^2_{01} = \frac{1}{M} \sum_{i=1}^{c} \left[\frac{(M f_{ii} - r_i s_i)^2}{r_i s_i} \right]$$

$$S\left(\sqrt{2\chi^2_{01}}\right) = \sqrt{2\chi^2_{01}} - \sqrt{2c-1}$$

2.3.2 The Chi-squared for Symmetry (or association).

$$\chi^2_{02} = \frac{1}{M} \sum_{\substack{i=1 \\ (i<j)}}^{c} \sum_{j=1}^{c} \left[\frac{\left[M(f_{ij}+f_{ji}) - (r_i s_j + r_j s_i) \right]^2}{(r_i s_j + r_j s_i)} \right]$$

$$S\left(\sqrt{2\chi^2_{02}}\right) = \sqrt{2\chi^2_{02}} - \sqrt{(c-1)(c-2)-1}$$

2.3.3 The Chi-squared for Bias.

$$\chi^2_{03} = M \sum_{\substack{i=1 \\ (i<j)}}^{c} \sum_{j=1}^{c} \left[\frac{(f_{ij} - f_{ji})^2}{r_i s_j + r_j s_i} \right]$$

$$S\left(\sqrt{2\chi^2_{03}}\right) = \sqrt{2\chi^2_{03}} - \sqrt{(c-1)(c-2)-3}$$

Note: $\chi^2_{01} + \chi^2_{02} + \chi^2_{03} = \chi^2$ of section 2.1.3

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

2.3.4 For the (selected) .1 Overall
.2 On-cut
.3 Off-cut
digraphic distributions compare the fit of the distribution of each diagonal with the expected distribution by means of the Chi-square and sigmage of Chi-square. (The print out will identify each Chi-Square and sigmage with the subscript of the category of the column in which the diagonal intersects the first row.)

$$\chi^2_{D_j} = \frac{1}{M} \sum_{i=1}^{c} \left[\frac{\left(M f_{i,(i+j) \bmod c} - r_i \cdot s_{(i+j) \bmod c} \right)^2}{r_i \cdot s_{(i+j) \bmod c}} \right]$$

$$S\left(\sqrt{2\chi^2_{D_j}}\right) = \sqrt{2\chi^2_{D_j}} - \sqrt{2c-1}$$

$$(j = 0, 1, 2, \cdots, (c-1).)$$

where D_j denotes the jth diagonal for $j = 0,1,2,\cdots,(c-1)$.

Example: The diagonal, D_1, for a 10 x 10 digraphic matrix includes $f_{1,2}; f_{2,3}; f_{3,4}; \cdots; f_{9,10}; f_{10,1}$.

2.4 Poisson printout of the digraphic distribution.

2.4.1 For the (selected) .1 Overall
.2 On-cut
.3 Off-cut
digraphic distributions print out the number of digraphs that occur each number of times: 0, 1, 2, $\cdots$, $(t-1)$, (t or more). The number t is the smallest integer equal to or greater than the expected count per cell plus 10 sigmas.

Does the PGM compute t ??

$$t \geq \frac{M + 10\sqrt{M(c^2-1)}}{c^2}$$

2.4.2 For the (selected) .1 Overall
.2 On-cut
.3 Off-cut
digraphic distributions print out the actual digraphs that occur m (or some other specified number) or more times, where m is equal to the expected count per cell plus 8 sigmas. For each such digraph, print the number of times it occurred.

$$m = \frac{M + 8\sqrt{M(c^2-1)}}{c^2}$$

DECLASSIFIED
NW55649
AUTHORITY
BY NARA DATE: 4/25

2.5 Digraphs within group.

For selected group sizes 2 to 6, count the digraphs formed by the characters in the positions selected from the following list of possibilities.

Gp. size 2:	1-2	2-1										
" 3:	1-2	1-3	2-3	3-1								
" 4:	1-2	1-3	1-4	2-3	2-4	3-4	4-1					
" 5:	1-2	1-3	1-4	1-5	2-3	2-4	2-5	3-4	3-5	4-5	5-1	
" 6:	1-2	1-3	1-4	1-5	1-6	2-3	2-4	2-5	2-6	3-4	3-5	
	3-6	4-5	4-6	5-6	6-1.							

(Note: 2-1, 3-1, 4-1, etc. indicate the last character in one group and the first character in the following group).

Perform for these digraphic distributions the subroutines selected from the list: Subroutines 2.5.1.1, 2.1.2, 2.1.3, 2.2.1, 2.2.2, 2.2.3, 2.3.1, 2.3.2, 2.3.3, 2.3.4, 2.4.1.

2.5.1.1 Print out the digraphic distribution of the selected counts of digraphs within group.

2.6 Digraphs at offsets (skip digraphs).

For each set of selected values of the parameters i, j, k count the digraphs formed as follows: the first digraph is formed by the character of the stream in position i as first element and by that in position j as second element. The parameter k, the skip, is added to both i and j successively to locate the positions of the elements of successive digraphs. Note that the sponsor may select more than one set of integers i, j, and k. This process of forming digraphs continues until the stream is exhausted. Perform for these digraphic distributions the subroutines selected from the list: Subroutines 2.6.1.1, 2.1.2, 2.1.3, 2.2.1, 2.2.2, 2.2.3, 2.3.1, 2.3.2, 2.3.3, 2.3.4, 2.4.1.

Example: If i = 1, j = 5, k = 3, then from the stream

A B C D E F G H I J K L M N O P Q R S T U . . . . .

the digraph count will consist of

AE, DH, GK, JN, MQ, PT, ...

DECLASSIFIED
NW 55649
AUTHORITY
BY LJ NARA DATE: 4/25

2.6.0.1 Combined Skip Digraph Distribution

Collect into one digraphic matrix the sum of the counts for all distributions corresponding to selected sets of parameters i, j, k. Perform for this distribution the subroutines selected from the list: Subroutines 2.6.1.1, 2.1.2, 2.1.3, 2.2.1, 2.2.2, 2.2.3, 2.3.1, 2.3.2, 2.3.3, 2.3.4, 2.4.1.

2.6.1.1 Print out the selected digraphic distributions.

Example: For the set of parameters

(i, j, k) = (1,4,3), (2,5,3), and (3,6,3)

the combined distribution, for the stream in the last example, will consist of the digraphs

AD, DG, GJ, JM, MP, PS. . . . . .

BE, EH, HK, KN, NQ, QT, ... CF, FI,

IL, LO, OR, RU . . . . . .

DECLASSIFIED
NW55649
AUTHORITY
BY NARA DATE: 4/25

overlapping ?

3. MONOGRAPHIC LOCAL ROUGHNESS

3.1 For each offset selected in the range 1 to 33 print the offset, the total number of monographic coincidences, the expected number of coincidences, and the sigmage. If any sigmage exceeds 3.5 (or other specified value) print out the distribution of coincident letters for that offset.

$$S(H) = \frac{c\left(H - \frac{P\delta}{c}\right)}{\sqrt{P\delta\,(c-\delta)}}.$$

Expected value, $E(H) = \frac{P\delta}{c}$.

Where: H = total number of coincidences

P = number of valid comparisons.

δ = monographic delta I.C.

3.2 For selected offsets 2 to 6, print subtotals of monographic coincidences according to "position-in-group".

Example: For offset 3 and the stream A B C A F C B F E A J E

The subtotals are

Position-in-group	1:	1
Position-in-group	2:	1
Position-in-group	3:	2

3.3 For selected offsets of any size up to stream length minus c, print the offset, the total number of monographic coincidences and the sigmage. If any sigmage exceeds 3.5 (or other specified value) print out the distribution of coincident letters for the associated offset.

DECLASSIFIED
NW55649
AUTHORITY
BY NARA DATE: 4/25

4. POLYGRAPHIC REPEATS

4.1 For each size m-graph (polygraph m letters long) $m = 2,3,\ldots,p$ (where p is determined below) print the total number of valid m-graphs, ${}_mN$, and the delta IC and sigmage of the distribution.

p is the smallest integer in the set $m = 3,4,5,\ldots$ such that

$$\frac{{}_mN({}_mN-1)}{c^m} \leq 40.$$

$$\delta_m = \frac{c^m}{{}_mN({}_mN-1)}\left[\left(\sum_{i=1}^{c^m} {}_mf_i^2\right) - {}_mN\right]$$

$$S(\delta_m) = \frac{{}_mN(\delta_m - \delta_{m-1}\cdot\delta_1)}{\sqrt{2(c^m-1)}}$$

Where: ${}_mf_i$ = frequency of m-graph i $(i = 1,2,\ldots,c^m)$.

${}_mN$ = total number of valid m-graphs.

δ_1 = monographic delta I.C.

4.2 For each size m- graph $(m = 2,3,\ldots,p)$ show the index of cohesion.

$$i_m = \frac{\delta_m}{\delta_{m-1}\cdot\delta_1}.$$

4.3 For the p determined in 4.1 print the following information for all repeated p-graphs or larger polygraphs. For each repeat print the position of the first character of its first occurrence and the offset to the first character of its second occurrence. Print the length of the repeat and print the repeat itself.

If for any integer m in the sequence $m = 3,4,5,\ldots,p$ the sigmage $S(\delta_m)$, computed in 4.1, is equal to or greater than + 3.5 (or other specified value), then print the same information as in the first paragraph of this section for all repeated m-graphs or longer polygraphs. (4.3)

Note: Print polygraphs in their longest form. Thus a hexagraph is not to be considered as two consecutive pentagraphs, etc...

DECLASSIFIED
NW556495
AUTHORITY
BY LJ NARA DATE: 4/25

4.4 Sum Check

Print out, for each polygraphic repeat of size (p plus k) or longer, the sum (modulo c) of those characters of the stream starting with the first character of the first occurrence of the polygraph and ending with the last character before the second occurrence. Any dit as part of the sum shall yield a final sum of dit. (The number k is a non-negative integer to be selected by sponsor).

Example: For the polygraphic repeat 523 shown below

...0 7 8 5 2 3 5 7 1 8 6 4 5 2 3 9 6 1 7 ...,

the sum is

5+2+3+5+7+1+8+6+4 = 1 (modulo 10).

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

5. WIDTH WRITEOUTS

5.1 For each width in a range of widths selected (from 2 to 128) and for each other width selected (over 128), print the width size w, the average columnar delta I.C. and the root mean square sigmage.

$$\overline{{}_w\delta} = \sum_{k=1}^{w} \frac{N_k \cdot {}_w\delta_k}{N}$$

$$\overline{S} = \sqrt{\frac{\sum_{k=1}^{w} [S({}_w\delta_k)]^2}{w}}$$

Where: N_k = total number of valid characters in column k.

$N = \sum_{k=1}^{w} N_k$ = total number of valid characters in the stream

${}_w\delta_k$ = the delta I.C. for the kth column of stream written on a width of w.

5.2 For each column of each width selected compute the monographic delta I.C. and its sigmage. For those columns with sigmage greater than + 3.0 (or other specified value), print the width size w, the column number, the delta I.C., and its sigmage.

$${}_w\delta_k = \frac{c}{N_k(N_k-1)}\left[\left(\sum_{i=1}^{c} f_{ik}^2\right) - N_k\right] \quad , \quad k = 1, 2, \ldots, w.$$

$$S({}_w\delta_k) = \frac{N_k({}_w\delta_k - \delta)}{\sqrt{2(c-1)}} .$$

Where f_{ik} = frequency of category i in column k.

DECLASSIFIED
NW55649
AUTHORITY
BY NARA DATE: 4/25

5.3 For each selected set of parameters i, j ; w specifying a pair of columns i, j of the stream written on width w, compare the distributions of the columns by means of the Chi-square and sigmage of chi-square of section 1.4.4, and print the parameters and these statistics.

5.4 For each selected set of parameters s; i, j ; w specifying a pair of columns i, j of the stream written on width w, and a slide s to be applied to column j, compare the distribution of column i against that of column j with slide s by means of the Chi-square and sigmage of Chi-square of section 1.4.5, and print the parameters and these statistics. In this subroutine for a given width w and pair of columns i, j, one designates the slides s that are to be run.

DECLASSIFIED
NW55645
AUTHORITY
BY LJ NARA DATE: 4/25

8. DERIVED STREAMS

From the stream(s) submitted by sponsor, derive selected streams of the types included below, or other types requested by sponsor, and then apply SUPERSTETH routines (selected from 1. to 5.) to such derived streams.

Provide for print out of any derived stream.

8.1 Delta Streams

8.1.1 Form the delta stream Δ_m for a selected integer m [$1 \leq m \leq$ (input stream length minus C)]. If the stream is $S_1, S_2, S_3, \ldots, S_N$ then the Δ_m stream is $(S_{m+1} - S_1), (S_{m+2} - S_2), \ldots, (S_N - S_{N-m})$. The subtraction will be taken modulo C (or any other chosen modulus up to 64). Use $\emptyset$ for the first character of the alphabet, 1 for the second, ..., and (C-1) for the last character of the alphabet. The input stream can be the original stream or a previously derived stream.

8.1.2 Same as 8.1.1 but reverse the direction of the subtractions, $(S_1 - S_{m+1})$, etc.

Note: If the modulus is larger than C, assign symbols to the newly-introduced characters from (C+1) up to modulus size. Print a list of the assignments made. Select from the unused symbols in the set

A, B, . ., Z , $\emptyset$, 1, . . ., 9, 37-, 38-, ...,

64-.

8.2 Decimation

8.2.1 For decimation m, obtain any selection of up to m substreams ${}_i d_m$ ($i = 1, 2, \ldots, m$). Any such substream may then be put through the SUPERSTETH routines.

($2 \leq m \leq \frac{\text{stream length}}{2}$).

Example: $m = 3$

${}_1d_3$: $S_1, S_4, S_7, S_{10}, \ldots$

${}_2d_3$: $S_2, S_5, S_8, S_{11}, \ldots$

${}_3d_3$: $S_3, S_6, S_9, S_{12}, \ldots$.

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

8.2.2 Combine a selection of the m substreams (in any prescribed order) into one long stream.

Example: Combining the substreams of the previous example in 1, 3, 2 order, we get the stream

$$S_1, S_4, S_7, \ldots S_3, S_6, S_9, \ldots S_2, S_5, S_8, \ldots .$$

8.3 De-Fibonaccied Streams.

8.3.1 Using the rule of combination (modulo C):
$[(S_1 + S_j) - S_k]$, $[(S_2 + S_{j+1}) - S_{k+1}]$, $[(S_3 + S_{j+2}) - S_{k+2}]$, . . .
derive a de-Fibonaccied stream, where j may be selected from 1 to 51 and k from 2 to 101 (always $j<k$).

8.4 Teleprinter Arithmetic Streams

8.4.1 Using sponsor's coding, obtain "delta-n" streams at selected interval m ($1 \leq m \leq (N-2)$) from level by level combination of stream characters:
(S_1, S_{m+1}), (S_2, S_{m+2}), . . . , under the rule of combination selected from the list of ten below:

	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Rule 6
	0 1	0 1	0 1	0 1	0 1	0 1
0	0 1	1 0	0 1	0 0	0 0	0 1
1	1 0	0 1	0 0	1 0	1 0	1 1

Rules 7, 8, 9, 10 are the complements of Rules 3, 4, 5, 6, respectively. (Note that Rule 2 is the complement of Rule 1.)

8.4.2 Convert the stream of characters into a single stream of bits by reading all levels of the first character, then all levels of the second character, then all levels of each succeeding character. In every case, read levels within a character in a sequence selected by the sponsor.

EXAMPLE: The stream of 6-level characters

~~0 1 1~~
1 0 1
0 0 1
0 1 0 0
1 1 0
0 0 0

read in 1 2 3 4 5 6 order becomes

Ø 1 0 0 1 0 / 0 0 1 1 0 / 1 1 0 0 0.

When read in 3 5 2 1 4 6 order it becomes

0 1 1 0 0 0 0 1 0 1 1 0 1 0 1 1 0 0.

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

SECRET

8.4.3 - transposition of levels

8.5 Substitution streams.

8.5.1 Derive a new stream by substitution from a given alphabet of an original stream to a new alphabet, using a selected type of substitution, as specified by the sponsor. A list of some suggested types is included below.

Type 1. Substitution n to 1 (where $1 < n < \frac{9}{2}$ for at least one character of the original alphabet).

EXAMPLE. Substitution table:

0	1	2	3	4	5	6	7	8	9
A	B	C	D	E	F	G	H	I	J
K	L	M	N	O	P	Q	R	S	T
U	V	W	X	Y	Z				

In the derived stream: A, K, U become Ø
B, L, V become 1, etc.

Type 2. Substitutions 1 for 1 or 2 for 1.

EXAMPLE. Substitution table

	3	4	5	6	7	8	9
	A	B	C	D	E	F	G
0	H	I	J	K	L	M	N
1	O	P	Q	R	S	T	U
2	V	W	X	Y	Z		

In the derived stream: A through G convert to mononomes (e.g A becomes 3).
H through Z convert to dinomes (e.g. H becomes 03, Z becomes 27).

Type 3. Substitutions 2 for 1

EXAMPLE. Substitution table

	5	6	7	8	9
0	A	B	C	D	E
1	F	G	H	I	K
2	L	M	N	O	P
3	Q	R	S	T	U
4	V	W	X	Y	Z

In the derived stream each letter converts to a dinome.

A becomes 05, B becomes 06, ---, Z becomes 49.

SECRET

DECLASSIFIED
NW5564G
AUTHORITY
BY LJ NARA DATE: 4/25

Type 4. Substitution 1 for 2.

EXAMPLE. Using the same table as in the previous example, convert 05 to A, 06 to B, ..., 49 to Z.

Provide for substitutions of other types specified by sponsor, using substitution matrices up to size c by c.

8.6 Allow for derivation of other streams by similarly simple rules to be specified by sponsor.

DECLASSIFIED
NW5561+9
AUTHORITY
BY LJ NARA DATE: 4/25

9. COMBINED STREAMS

Make provisions so that derived streams (or sponsor-submitted streams) can be combined into one long stream which will then be put through selected SUPERSTETH routines.

Place such sets of substreams, in some prescribed order, into the long stream at successive fixed intervals j $(100 \leq j \leq 2000)$ filling in the gaps between substreams with dits. The length of each substream should, of course, be not greater than the interval selected.

For example, for interval 500, the first substream will occupy positions 00001 to L_1, (its length), the second will occupy 00501 to (00500 plus L_2), etc. Each gap, of length $(500 - L_i)$, between the i th substream and the succeeding one, should be filled in with dits. $(i = 1, 2, 3, \ldots)$.

DECLASSIFIED
NW55649
AUTHORITY
BY LJ NARA DATE: 4/25

APPENDIX A

SUPERSTETH REBUS (Request Form)

PREPARED BY: MPRO-503

The Supersteth Rebus (17 pages) is not included in this reprint. Copies of the Rebus may be obtained from PROD-03 or MPRO-05.

Dear Patron:

We regret that the enclosed photocopies are the best we were able to obtain using our normal reproduction process. This is caused primarily by the age and faded conditions of some of the documents from which these copies are made.

BEST AVAILABLE COPY